Estimación de la pose humana 2D en imágenes estéreo

Manuel Ignacio López Quintero

Directores:
Manuel Jesús Marín Jiménez
Rafael Muñoz Salinas

Programa de Doctorado: Ingeniería y Tecnología
Línea de investigación: Sistemas Inteligentes en Visión
Departamento de Informática y Análisis numérico
Universidad de Córdoba

13 de junio de 2016

TITULO: *ESTIMACIÓN DE LA POSE HUMANA 2D EN IMÁGENES ESTÉREO*

AUTOR: *Manuel Ignacio López Quintero*

Campus de Rabanales
Ctra. Nacional IV, Km. 396 A
14071 Córdoba

www.uco.es/publicaciones
publicaciones@uco.es

TÍTULO DE LA TESIS: ESTIMACIÓN DE LA POSE HUMANA 2D EN IMÁGENES ESTÉREO

DOCTORANDO/A: Manuel Ignacio López Quintero

INFORME RAZONADO DEL/DE LOS DIRECTOR/ES DE LA TESIS

(se hará mención a la evolución y desarrollo de la tesis, así como a trabajos y publicaciones derivados de la misma).

El alumno ha realizado un excelente trabajo durante estos años en su labor investigadora. Su esfuerzo ha permitido realizar importantes contribuciones en el ámbito de la Visión por Ordenador, y en particular en el ámbito de la estimación de la pose humana utilizando múltiples puntos de vista con el uso de modelos pictóricos. Los trabajos realizados han permitido la publicación del trabajo:

Manuel I López-Quintero, Manuel J Marín-Jiménez, Rafael Muñoz-Salinas, Francisco J Madrid-Cuevas, Rafael Medina-Carnicer "Stereo Pictorial Structure for 2D articulated human pose estimation", Machine Vision and Applications, Volume 27, Issue 2, pp 157-174.

Además, como aportación adicional a la tesis, se ha realizado el trabajo

Manuel I Lopez-Quintero, Manuel J. Marín-Jiménez, Rafael Muñoz-Salinas, Rafael Medina-Carnicer, "Mixing Body-Parts for 2D Human Pose Estimation in Stereo Videos"

que se encuentra actualmente en proceso de revisión.

Por estas razones, entendemos que la Tesis doctoral presentada por Manuel Ignacio cumple con los requisitos necesarios de excelencia para otorgarle el grado de Doctor en Informática.

Por todo ello, se autoriza la presentación de la tesis doctoral.

Córdoba, 13 de junio de 2016

Firma del/de los director/es

Fdo.:Rafael Muñoz Salinas

Fdo.:Manuel Jesús Marín-Jiménez

Resumen

La Estimación de la Pose Humana es el proceso de obtener la configuración espacial de las partes del cuerpo en imágenes. Frente a los métodos monoculares, que recuperan la pose a partir de una sola imagen, los métodos estéreo usan un par de imágenes para realizar el proceso, siendo capaces de aprovechar la redundancia de información y así mejorar la precisión. Este trabajo de Tesis se centra en adaptar técnicas monoculares de estimación de la pose ya existentes para que sean capaces de aprovechar las ventajas del uso de información estéreo.

La primera contribución de esta tesis es una nueva técnica para estimar la pose 2D de personas en imágenes estéreo basado en una restricción de similitud que permite la colaboración entre dos estimadores de pose. Nuestra propuesta mejora la precisión de las poses estimadas en comparación con técnicas monoculares de estimación de la pose ejecutadas de forma independiente en cada vista de la imagen estéreo.

La segunda contribución es una base de datos para el problema de la estimación de la pose humana en imágenes estéreo. Para validar experimentalmente nuestras propuestas, hemos creado una nueva base de datos anotada de 630 imágenes estéreo que muestran personas en entornos diferentes, con ropa variada y diversa iluminación. La base de datos muestra a las personas en posición vertical con una gran variedad de poses de brazos que cubren todo el espacio de posibles configuraciones de poses.

La tercera contribución es un nuevo método para estimar la pose 2D de personas en secuencias de vídeo estéreo. El método comienza con una reducción de las posibles localizaciones de las partes del cuerpo usando información de color y de disparidad. A continuación se utiliza información *a priori* para la localización de las partes del cuerpo más estructuradas. Por último, un método de recombinación de partes del cuerpo se aplica en la secuencia estéreo para obtener la mejor configuración de las partes del cuerpo. Los experimentos demuestran que la propuesta consigue mejores resultados que el actual estado del arte.

Abstract

Human Pose Estimation (HPE) is the task of obtaining the spatial configuration of human body parts from images. Methods recovering the human pose from a single image are called monocular approaches while those using image pairs are called stereo approaches. Stereo images provide extra information that can be employed to improve the results obtained by monocular approaches.

This Thesis considers the problem of 2D human pose estimation on stereo images. To this end, three contributions are provided.

The first contribution of this thesis is a new technique to automatically detect and estimate the 2D pose of humans in stereo images. The proposed method is based on a similarity constraint that promotes a collaboration between two pose estimators. We show experimentally that our proposal improves the accuracy of the estimated poses when compared to standard HPE techniques running independently on each image.

The second contribution is a dataset for the problem of human pose estimation in stereo image. To experimentally validate our approach, we have created a new annotated dataset of 630 stereo image from stereo videos depicting people in different backgrounds, clothing, lighting or locations in the image frames. The dataset contains upright people in a great variety of arms poses, covering the space of possible configurations quite uniformly.

The third contribution is a new method to estimate the 2D pose of humans in stereo videos sequences. The proposed pipeline starts by constraining the possible location of body joints by exploiting color and disparity information, and adding location priors to the most structured joints. Finally, a body limb recombination method is applied along the stereo sequence to obtain the best configuration of the body joints. The experiments show that our method obtains better average results than the state-of-the-art.

Declaración

Manuel Jesús Marín Jiménez y Rafael Muñoz Salinas, profesores del Departamento de Informática y Análisis Numérico de la Universidad de Córdoba, INFORMAN que la Tesis Doctoral realizada por D. Manuel Ignacio López Quintero con título *Estimación de la pose humana 2D en imágenes estéreo* ha sido desarrollada bajo su dirección y que la misma presenta contribuciones originales que permiten que la misma pueda ser defendida.

Director: *Director:*

Manuel Jesús Marín Jiménez Rafael Muñoz Salinas

Córdoba 10 de junio de 2016

A mi madre

Agradecimientos

En primer lugar agradecer a mis directores, los doctores Manuel Jesús Marín Jiménez y Rafael Muñoz Salinas, por toda la labor realizada durante esta Tesis Doctoral. Su ayuda y guía ha sido fundamental para la culminación de todo este trabajo.

También agradecer a Rafael Medina Carnicer y, en general, al equipo de profesores de AVA por ofrecerme la posibilidad de colaborar en su grupo y de trabajar con la financiación necesaria.

Quiero dar las gracias a mis compañeros de laboratorio: David López Fernández, Eusebio Jesús Aguilera Aguilera, Sergio Garrido Jurado y Víctor Manuel Mondéjar Guerra. Han sido tres años y dos meses de compartir horas y horas juntos todos los días. Su ayuda, compañerismo y amistad durante todo este tiempo ha permitido que la Tesis haya sido una etapa que recordaré con mucho cariño.

Por último agradecer a mi familia: mi madre, mi padre, mis hermanos Javi y Chechu, Titi, Pepe, Gema y, por supuesto, a Julio y a la pequeña Emma. Ellos han sido un apoyo indispensable en los buenos y malos momentos.

Agradezco a Tina porque siempre ha estado pendiente del estado de mi Tesis. Sus ánimos y alegría son algo que nunca olvidaré.

Y, para terminar, agradezco a mis amigos que siempre han estado ahí para sacarme la mejor de las sonrisas: Antonio, Carazo, Duro, Graci, Falele, Fernando, Fran, Grego, José, Lizana, Márquez, Miguel, Paco, Pedro, Peter, Ricardo, Rodo, Sergio y Wo.

Muy problamemente se me olvidará a alguien, pero estoy seguro que me lo perdonará.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

Las Ciencias de la Computación estudian los principios teóricos de la información así como su implementación y aplicación. Dentro de sus campos de estudio, la Visión Artificial se encarga de analizar, procesar y, sobre todo, entender la información contenida en imágenes digitales del mundo real. Para ello, las imágenes digitales son transformadas en información numérica y simbólica y los diferentes modelos se encargan de entender el contenido de la imagen. Los modelos no son generalistas, sino que suelen ser específicos para ciertas tareas. Por ejemplo, en el trabajo de Viola y Jones [1] se proponen modelos específicos para la detección de caras en tiempo real. Otro ejemplo es el trabajo de Felzenszwalb *et al.* [2], donde se definen modelos concretos para la detección de objetos en una imagen.

Existe un gran número de aplicaciones de Visión Artificial, algunos ejemplos son:

- **Metrología óptica 2D y 3D:** en el campo de la metrología, los sistemas de Visión Artificial obtienen las medidas físicas del objeto y comprueban que se corresponden con el patrón exigido [3].
- **Reconocimiento óptico de caracteres:** del inglés *Optical Character Recognition* (OCR), los sistemas de Visión Artificial en este área se encargan de la identificación de símbolos o caracteres en una imagen [4].
- **Sistemas de vigilancia:** el seguimiento de personas a través de una o varias cámaras es una de las aplicaciones de la Visión Artificial que más está creciendo en cuanto a uso debido al aumento de la seguridad en edificios [5].
- **Identificación de personas:** en la especialidad de Visión Artificial de biometría, la utilización de imágenes de caras, retinas de los ojos o, por ejemplo, huellas dactilares posibilita la identificación de individuos [6].

(a) **(b)**

Figura 1.1: ***Estimación de la pose humana 2D en una imagen****. Los métodos de estimación de la pose monoculares infieren la posición de las partes del cuerpo a partir de una sola imagen. (a) Imagen de una persona con encuadre en plano americano (cabeza hasta las rodillas). (b) Pose estimada usando el método monocular de Eichner et al. [10].*

- **Reconocimiento de acciones de personas:** identificar las poses humanas en un período de tiempo permite identificar las acciones llevadas a cabo [7].
- **Meteorología:** reconocer e interpretar imágenes meteorológicas para clasificar el tiempo es una aplicaciones más de la Visión Artificial [8].

Como consecuencia del gran número de aplicaciones, la Visión Artificial tiene una amplia diversidad de líneas o ramas de investigación. La dificultad de entender con exactitud imágenes digitales del mundo real hace que muchas de estas líneas tengan una gran cantidad de retos aún sin resolver. Una de ellas es la línea de investigación *estimación de la pose humana* que ha sido estudiada ampliamente durante los últimos 40 años [9].

La estimación de la pose humana (en inglés *Human Pose Estimation* o HPE) es el proceso de predecir las posiciones 2D y/o 3D de cada una de las partes del cuerpo de una o varias personas en imágenes o vídeos. Un vídeo es una sucesión ordenada de imágenes. Las imágenes pueden tener una vista (imágenes monoculares), dos vistas (imágenes estéreo) o varias vistas (imágenes multivista). Los métodos de estimación de pose humana monoculares usan como entrada imágenes o vídeos monoculares. Los métodos de estimación de pose humana estéreo usan imágenes o vídeos estéreo. Los métodos de estimación de la pose humana multivista, también denominados multicámara, usan imágenes o vídeos multivista. Un ejemplo de aplicación de un método de estimación de la pose 2D monocular en una imagen de color puede verse en la figura 1.1.

Como se ha dicho anteriormente, las propuestas de estimación de la pose humana estéreo (en inglés *Stereo Human Pose Estimation* o SHPE) utilizan imágenes o vídeos estéreo. Una imagen estéreo está compuesta por dos imágenes que corresponden a dos vistas de una misma escena. Si la imagen estéreo se ha obtenido mediante una cámara estéreo, una de las vistas corresponde a la lente izquierda y

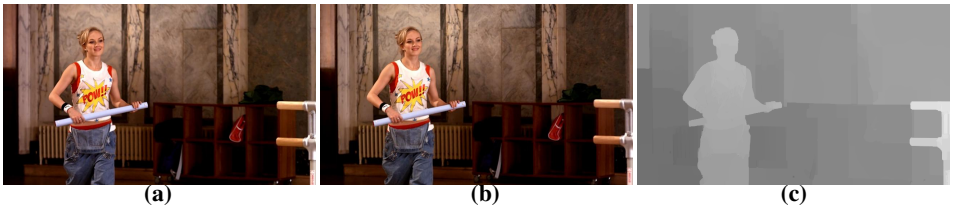


Figura 1.2: ***Imagen estéreo****. Una imagen estéreo está compuesta por dos imágenes que han sido obtenidas en un mismo instante de tiempo pero en posiciones ligeramente diferentes. A cada una de estas imágenes se les llama vista izquierda o derecha dependiendo de su posicionamiento con respecto al origen fijado. (a) Vista izquierda de una imagen estéreo. (b) Vista derecha de una imagen estéreo. (c) Mapa de disparidad calculado usando el método de Ayvaci et al. [11].*

la otra a la lente derecha. Una de las características de una imagen estéreo es que puede calcularse su imagen o mapa de disparidad. La disparidad en una imagen estéreo es la diferencia relativa entre la posición de los puntos de una vista y la posición correspondiente a dichos puntos en la otra vista. Un ejemplo de una imagen estéreo y su mapa de disparidad puede apreciarse en la Figura 1.2. Para el mapa de disparidad de este ejemplo, los puntos más claros corresponden a posiciones más cercanas a la cámara y los puntos más oscuros, sin embargo, a posiciones más distantes.

Los métodos de estimación de la pose humana estéreo se benefician de la información adicional que este tipo de imágenes posee, obteniendo así mejores resultados cualitativos y cuantitativos que los métodos de estimación monoculares. Véase, por ejemplo, la fila inferior de la Figura 1.3-a (HPE), donde un método monocular produce un error en cada una de las vistas de la imagen estéreo. Sin embargo, el método estéreo, fila inferior de la Figura 1.3-b (SHPE), estima las poses correctamente. A través de este ejemplo se puede observar que los métodos monoculares producen una estimación de la pose diferente en cada vista. En cambio, algunos métodos estéreo como el de este ejemplo, al combinar la información de la vista izquierda y derecha producen una configuración de la pose común a ambas vistas, pudiendo corregir los errores que aparecen por separado.

La estimación de la pose humana también se puede obtener con equipos de captura de movimiento (en inglés *motion capture* o, en su abreviación, *mocap*). En este caso, una persona posee en su cuerpo marcadores físicos (en inglés *markers*) y con varias cámaras se infiere la pose, frecuentemente 3D, de la persona. Aunque la estimación de la pose es muy precisa, el equipo necesario para realizarlo es muy costoso. Además, este tipo de estimación generalmente se realiza en lugares o espacios muy preparados para ello siendo obligatorio que la persona disponga de marcadores en su cuerpo. Teniendo en cuenta lo expuesto anteriormente, en esta tesis los métodos de predicción de la pose serán sin ningún tipo de marcadores en

Figura 1.3: ***Estimación de la pose monocular y estéreo****. (a) Poses estimadas por un modelo de estimación de pose monocular, en este caso [10], donde cada vista es procesada independientemente proporcionando poses diferentes. (b) Poses estimadas con el método SHPE propuesto en el Capítulo 3. Como se puede observar, una pose común es estimada en la imagen estéreo obteniendo la misma configuración de las partes del cuerpo para cada una de las vistas.*

las personas (*markerless*).

1.1. Motivación y Objetivos

Dentro del marco general de la estimación de la pose humana sin marcadores, la presente Tesis tiene como objetivo demostrar que el uso de información estéreo permite obtener mejoras significativas respecto al uso de información monocular. Por ello, nuestro trabajo se centrará en proponer nuevas técnicas estéreo basándonos en técnicas monoculares ya existentes las cuales podrán beneficiarse de las ventajas del uso de información estéreo.

Entre las técnicas existentes para la estimación de la pose, aquellas basadas en modelos pictóricos [12] son las que tienen un mayor auge debido a su fácil interpretación e implementación, además con ellas se obtienen actualmente buenos resultados [13]. Los modelos pictóricos son modelos gráficos probabilísticos donde las partes del cuerpo de una persona están representadas por un campo aleatorio condicional [14] y permiten estimar la configuración espacial de las partes de un objeto, en nuestro caso, la persona.

Si bien la mayoría de las técnicas desarrolladas de estimación de la pose humana se basan en el uso de información monocular, numerosos estudios previos en diversos ámbitos muestran que el uso de información estéreo permite mejorar los

procesos, como por ejemplo en detección y seguimiento de personas [15], navegación [16], análisis biométrico [17], etc. Pese a ello, pocos autores han tratado de aplicar estas ideas al problema de estimación de la pose con modelos pictóricos.

El primer primer objetivo de la Tesis será crear una base de datos estéreo que permita evaluar los diferentes métodos que vamos a desarrollar. Para ello, obtendremos secuencias de vídeo estéreo de internet, anotaremos las poses de las personas además de otros añadidos como regiones espaciales de detección de personas, transformaciones del plano proyectivas, etc.

En segundo lugar, se realizará una extensión de la popular técnica pictórica monocular de Eichner *et al.* [10] para que pueda hacer uso de la información estéreo. Esta técnica se basa en una reducción progresiva del espacio de búsqueda de las partes del cuerpo para aumentar las posibilidades de la estimación 2D de la pose.

Si bien en el objetivo anterior se desarrolla una técnica que permite hacer uso de información estéreo, aún no se explota de forma adecuada todo el potencial disponible en la base de datos desarrollada. Nos referimos al uso de la información temporal disponible en las secuencias de vídeo. A este respecto, existen algunas aportaciones para la estimación de la pose en secuencias de vídeo usando modelos pictóricos [18]. Por ello, nuestro ultimo objetivo, es extender la técnica de Cherian *et al.* [19] para secuencias de vídeo estéreo en lugar de secuencias de vídeo monoculares.

1.2. Retos

Entre todos los retos a enfrentar en esta Tesis, estos son los que se consideran más desafiantes:

- **Diferentes puntos de vista:** debido a la variabilidad de la posición de la cámara, existe una manifiesta dificultad al estimar la pose de personas observadas desde diferentes puntos de vista. Esto es debido a que la proyección 2D de un objeto articulado 3D causa, en determinadas circunstancias, incertidumbre en la pose.

- **Apariencia de la persona:** existe una gran variedad en cuanto a la apariencia de cada persona. Ya no sólo el físico propio de cada uno, sino también su vestimenta influye cuando se quiere estimar la pose. Abordar este problema de una manera exitosa es uno de los retos a alcanzar.

- **Ambigüedad de la pose:** la diversidad de poses en un ser humano es amplia. Las personas pueden estar en una gran cantidad de configuraciones diferentes. Si a ello sumamos la perspectiva y que los brazos pueden tener el mismo color que el torso por la ropa, el reto es aún mayor.

- **Cambios de escala:** en una imagen la persona está a una cierta distancia con respecto a la cámara. Esta distancia hace que la persona esté a una escala determinada. Los métodos propuestos en esta Tesis están preparados para estimar la pose independientemente de la escala de la persona. Esto supone un gran reto ya que otros métodos establecen una misma escala para todas las personas.

- **Efectos de compresión de vídeo, desenfoque y resolución:** las imágenes se obtienen a partir de fotografías hechas con una cámara pero también se pueden obtener extrayendo imágenes en un vídeo. Debido al gran tamaño en *bytes* de un vídeo, éste necesita comprimirse. Como consecuencia la calidad de las imágenes no será tan alta como en una fotografía. Además de los efectos de compresión de vídeo, la borrosidad y el desenfoque asociado a un movimiento rápido de la persona o de la cámara disminuye aún más la definición de la persona en la imagen. A todo esto se suma la resolución en píxeles de una imagen que depende de las características y propiedades de la cámara.

- **Oclusiones y auto-oclusiones:** para el problema de estimación de la pose humana, se dice que una parte del cuerpo está ocluida si no es visible desde el punto de vista actual de la cámara. También, dependiendo de la posición de la cámara, se pueden producir auto-oclusiones donde las mismas partes del cuerpo de una persona tapan otras partes de su cuerpo. Por ejemplo, una persona de perfil puede tapar un brazo completo o una persona de frente con los brazos completamente flexionados también puede ocasionar una auto-oclusión.

- **Iluminación y fondo:** otro de los problemas cuando se estima la pose es la iluminación y fondo en una escena. Si el fondo de una imagen tiene colores similares a la piel o a la ropa de la persona, el problema de la estimación aumenta. A esto se le añade la iluminación en la escena que puede variar la apariencia de una persona y el fondo a lo largo de una secuencia de vídeo.

Algunas de las dificultades anteriormente mencionadas pueden visualizarse en las imágenes de ejemplo de la Figura 1.4.

1.3. Contribuciones

Esta Tesis Doctoral presenta tres contribuciones principales.

- **Estimación de la Pose Humana Estéreo:** la primera de las contribuciones es un modelo de estimación de la pose humana para imágenes estéreo

Figura 1.4: ***Retos en la estimación de la pose humana****. Aquí se muestran algunas imágenes de ejemplo donde se visualizan los retos a afrontar. (a) Cambios de escala. (b) Efectos de emborronamiento debido al movimiento. (c) Oclusiones y auto-oclusiones. (d) Iluminación y fondo; parte izquierda: inicio de la secuencia de vídeo; parte derecha: final de la secuencia de vídeo.*

de nombre Estimación de la Pose Humana Estéreo (en inglés *Stereo Human Pose Estimation* o SHPE). Basándose en el trabajo de Eichner *et al.* [10], se propone una extensión de este método en cada una de las etapas así como un nuevo modelo para imágenes estéreo: Modelo Pictórico Estéreo. El método propuesto se encuentra disponible en [20].

Las pruebas experimentales revelan que SHPE mejora a otras propuestas del estado del arte como el trabajo de Eichner *et al.* [10], otros populares como Yang y Ramanan [21] e incluso métodos de coestimación de la pose como el de Eichner y Ferrari [22].

Este trabajo ha sido publicado bajo el titulado *Stereo Pictorial Structure for 2D Articulated Human Pose Estimation* en la revista Machine Vision and Applications.

Toda la información sobre esta contribución se encuentra en el Capítulo 3.

- **The Stereo Human Pose Estimation Dataset:** la segunda de contribución principal es una base de datos para estimación de la pose humana en imágenes estéreo titulado en inglés *The Stereo Human Pose Estimation Dataset* o SHPED.

  Esta base de datos dispone de 630 imágenes estéreo estructuradas en 42 secuencias de vídeo estéreo con 15 imágenes estéreo cada secuencia. Estas secuencias han sido extraídas de 26 vídeos estéreo de YouTube (`youtube.com`).

  SHPED está disponible para descargar en [23]. En dicha web se encuentra los archivos para obtener todas las imágenes estéreo, anotaciones, identificadores vídeos YouTube, transformaciones del plano proyectivas y ficheros de demostración.

  Toda la información sobre SHPED se encuentra en el Apéndice A.

- **Modelo de Recombinación de Partes Estéreo:** la tercera contribución principal es la de proporcionar un modelo denominado Modelo de Recombinación de Partes Estéreo para estimar la pose humana en secuencias de vídeo estéreo. A diferencia con la primera contribución donde se estima la pose sólo en imágenes estéreo, aquí se estima en secuencias estéreo donde se beneficia de la información temporal para obtener resultados con más precisión. Este método tiene como base el método de Cherian *et al.* [24] que trabaja con secuencias de vídeo monoculares.

  Las pruebas experimentales revelan que el Modelo de Recombinación de Partes Estéreo gana a otras propuestas del estado del arte como el trabajo de Cherian *et al.* [24], el popular trabajo de Yang y Ramanan [21] e incluso los últimos métodos que usan redes neuronales convolucionales como el de Pfister [25].

  Como resultado de esta contribución se ha escrito el artículo *Mixing Body-Parts for 2D Human Pose Estimation in Stereo Videos* que ha sido enviado a una revista indexada en *Journal Citation Reports* (JCR).

  Toda la información sobre esta contribución se encuentra en el Capítulo 4.

1.4. Organización de la Tesis

Este documento está estructurado en un total de cinco capítulos, dos apéndices y una bibliografía.

En el Capítulo 1 se han explicado los objetivos, motivaciones, retos y contribuciones de esta Tesis Doctoral.

En el Capítulo 2 se realiza una revisión bibliográfica y de métodos del estado del arte. Los diferentes trabajos se categorizan según el número de puntos de vista o cámaras empleados: aproximaciones monoculares, aproximaciones estéreo y aproximaciones multicámara. También se esquematizan los métodos mediante una tabla.

En el Capítulo 3 se detalla el Modelo Pictórico Estéreo para la estimación de la pose humana en imágenes estéreo. Con una introducción a las estructuras pictóricas y al método en que se basa, Eichner *et al.* [10], se procede con las propuestas para expandir cada etapa. Por último, se muestran resultados experimentales de la propuesta que supera otros métodos del estado del arte monoculares y de coestimación de la pose.

En el Capítulo 4 se explica el Modelo de Recombinación de Partes Estéreo para la estimación de la pose humana en secuencias de imágenes estéreo. La propuesta se basa en Mezcla de secuencias de partes del cuerpo de Cherian *et al.* [19] y en ella se plantea una reducción del espacio de búsqueda así como un modelo con

dos procedimientos: Recombinación estéreo de extremidades y Extremidad mejor puntuada.

En el Capítulo 5 se concluye con un resumen así como una breve reseña de las publicaciones relacionadas. También, por último, se abren tres ramas principales de trabajo futuro para continuar con el proceso de investigación.

En el Apéndice A se muestran las tres bases de datos para estimación de la pose humana utilizadas en los experimentos. La base de datos Stereo Human Pose Estimation Dataset (ver A.1) se ha realizado para esta Tesis Doctoral.

En el Apéndice B se explican las tres métricas de evaluación utilizadas en los experimentos para medir los resultados. Para cada una de las métricas se explican las ventajas y desventajas de usar dicha medida de evaluación.

Por último, se lista la Bibliografía consultada a lo largo de la elaboración de este documento.

Capítulo 2

Revisión bibliográfica y métodos

Existen diferentes clasificaciones de los métodos del estado del arte sobre estimación de pose humana, uno de ellos es el de Liu *et al.* [26] que divide los métodos en cuatro categorías: estimación de pose 2D unipersona en imágenes, estimación de poses 2D multipersona en imágenes, estimación de pose 2D unipersona en vídeos y estimación unipersona de pose 3D en imágenes o vídeos. En esta tesis se ha optado por categorizar los diferentes trabajos (ver Tabla 2.1) según el número de cámaras o puntos de vista utilizados.

Aproximaciones monoculares: los métodos que abarca esta categoría son las propuestas que estiman la pose humana, tanto en 2D como en 3D, a partir de imágenes monoculares. Una imagen monocular es aquella imagen de una sola vista. Estas imágenes son capturadas habitualmente mediante cámaras de una sola lente o con cámaras con sensores de profundidad. Las primeras devuelven imágenes 2D de color y las segundas devuelven mapas de profundidad donde la información de cada píxel es la profundidad en metros con origen en dicha cámara.

Aproximaciones estéreo: esta categoría de aproximaciones estiman la pose a partir de imágenes estéreo. Una imagen estéreo utiliza dos imágenes monoculares tomadas desde dos posiciones diferentes. A estas dos imágenes monoculares frecuentemente se le llaman vista derecha y vista izquierda. Una imagen estéreo puede obtenerse de dos formas: mediante una cámara de una sola lente donde se toma una imagen monocular y después se desplaza dicha cámara [55] o mediante una cámara estéreo que es la forma más popular. Una cámara estéreo es un dispositivo con dos lentes cuya posición de dichas lentes están a la misma altura con

Tabla 2.1: ***Diferentes métodos para la estimación de la pose humana.*** *Abreviaturas - FE: Formato de Entrada, IC: Imagen de Color, MP: Mapa de Profundidad; TE: Tipo de Entrada, E: Estática, S: Secuencial; EP: Estimación de la Pose (2D o 3D); CP: Cantidad de Personas que puede estimar el método en una entrada, U: Unipersona, M: Multipersona*

Métodos	Vistas	FE	TE	EP	CP	Técnica principal
Lee y Cohen [27]	Mono	IC	E	3D	U	*Proposal maps*
Mori *et al.* [28]	Mono	IC	E	2D	U	*Normalized Cuts*
Felzenszwalb *et al.* [12]	Mono	IC	E	2D	U	Estructuras pictóricas
Ramanan [29]	Mono	IC	E	2D	U	Análisis iterativo
Ferrari *et al.* [18]	Mono	IC	S	2D	M	Reducción del espacio de búsqueda
Eichner y Ferrari [30]	Mono	IC	E	2D	M	Localizaciones del cuerpo *a priori*
Eichner *et al.* [10]	Mono	IC	E	2D	M	*A prioris* de orientación
Zuffi *et al.* [31]	Mono	IC	E	2D	U	Estructuras deformables
Yang y Ramanan [21]	Mono	IC	E	2D	M	Mezcla de partes del cuerpo flexibles
Cherian *et al.* [19]	Mono	IC	S	2D	U	Mezcla de secuencias de partes del cuerpo
Eichner y Ferrari [32]	Mono	IC	E	2D	M	Predictor de oclusión
Eichner y Ferrari [22]	Mono	IC	E	2D	M	Modelo de coestimación
Tompson *et al.* [33]	Mono	IC	E	2D	U	Híbrido de CNN y Markov
Pfister *et al.* [25]	Mono	IC	S	2D	U	Híbrido de CNN y flujos ópticos
Hernández *et al.* [34]	Mono	IC	E	2D	U	*Poselets*
Zhu y Fujimura [35]	Mono	MP	S	3D	U	Optimización con restricciones
Shotton *et al.* [36]	Mono	MP	E	3D	M	Bosque aleatorio seguido de *pooling*
Baak *et al.* [37]	Mono	MP	S	3D	U	Híbrido de optimización local con global
Ye *et al.* [38]	Mono	MP	E	3D	U	Híbrido de detección y refinamiento pose
Schwarz *et al.* [39]	Mono	MP	S	3D	U	Distancias geodésicas y flujos ópticos
Sun *et al.* [40]	Mono	MP	S	3D	U	Bosque de regresión condicional
Pons-Moll *et al.* [41]	Mono	MP	E	3D	U	MSIG
Jiu *et al.* [42]	Mono	MP	E	3D	U	CNN con bosque aleatorio de decisión
Guo y Qian [43]	Estéreo	IC	S	3D	U	Mezcla bayesiana experta
Yang y Lee [44]	Estéreo	IC	S	3D	U	Aprendizaje de arriba hacia abajo
Thang *et al.* [45]	Estéreo	IC	S	3D	U	EM con iteraciones de dos pasos
Sheasby *et al.* [46]	Estéreo	IC	S	3D	U	Optimización con doble descomposición
Lallemand *et al.* [47]	Estéreo	IC	S	3D	U	Descriptor de formas basado en rejilla
Seguin *et al.* [48]	Estéreo	IC	E	2D	M	Segmentación de articulación aprendida
Yao *et al.* [49]	Multi	IC	S	3D	U	Refinamiento de pose y acción estimada
Shen *et al.* [50]	Multi	IC	S	3D	U	Método de ajuste de plantilla
Sigal *et al.* [51]	Multi	IC	S	3D	U	Modelo *loose-limbed*
Amin *et al.* [52]	Multi	IC	S	3D	U	Estructura pictórica multivista
Burenius *et al.* [53]	Multi	IC	S	3D	U	Estructura pictórica 3D
Kazemi *et al.* [54]	Multi	IC	S	3D	U	Verosimilitudes 2D llevadas a 3D

respecto al nivel del suelo y separadas entre sí por unos pocos centímetros. Los métodos de estimación de la pose mediante cámaras estéreo son muy escasos pero en los últimos años está adquiriendo popularidad debido a los buenos resultados y al abaratamiento de cámaras estéreo. Uno de los objetivos primordiales de la tesis es precisamente demostrar las ventajas de usar imágenes estéreo con respecto a imágenes monoculares a la hora de estimar la pose humana.

Aproximaciones multicámara: este tipo de métodos calculan la pose humana en tres o más vistas. La ventaja de estas aproximaciones es que la estimación de la pose es más precisa que en aproximaciones estéreo y aún más que en aproximaciones monoculares. Las aproximaciones multicámara generalmente estiman también la pose 3D de la persona debido ya que hay más información disponible. Las múltiples vistas permiten a los métodos corregir ambigüedades en la pose y reducir los errores en la estimación. A menudo estas imágenes o vistas se obtienen desde posiciones bastante más separadas que en una imagen estéreo. Esto significa que, a diferencia de las cámaras estéreo, para obtener tres o más vistas es necesario montar un escenario propicio para capturar dichas imágenes. Además estas aproximaciones en su mayoría necesitan calibrar las cámaras, conocer la posición relativa entre ellas y confirmar que todas las imágenes están sincronizadas.

2.1. Aproximaciones monoculares

Dependiendo del tipo de imágenes monoculares (imágenes de color o imágenes/mapas de profundidad) podemos identificar dos subcategorías de aproximaciones monoculares para estimar la pose:

- **Aproximaciones basadas en color:** este tipo de métodos estiman la pose a partir de la información de color de una imagen. Si bien hay casos en que la estimación es factible debido a la aparición de poses habituales (fácilmente entrenables y detectables) o ropa sencilla (por ejemplo camisetas con colores lisos y sin estampados), otras veces resulta un problema desafiante debido a factores como la iluminación, oclusiones de las partes del cuerpo, poses complejas o ropa extravagante.
- **Aproximaciones basadas en mapas de profundidad:** estas aproximaciones estiman la pose a partir de una imagen o mapa de profundidad. El problema de la reconstrucción 3D de movimientos humanos complejos a partir de imágenes 2D de color es un problema difícil y a veces intratable. Este problema se hace más factible utilizando los mapas de profundidad monoculares según la información devuelta por una cámara de profundidad. Sin embargo, debido a la baja resolución y al ruido de este tipo de cámaras, además de producirse autooclusiones en los movimientos, la tarea de estimación de pose aún está lejos de ser simple.

2.1.1. Aproximaciones basadas en color

La estimación de pose humana en imágenes estáticas es un reto debido a la alta diversidad en las imágenes. Una aproximación que usa un método de cadenas de Markov Monte Carlo para estimar la pose 3D de la parte superior del cuerpo

humano es el trabajo de Lee y Cohen [27]. Ellos proponen un modelo generativo que comprime la estructura y la forma de las articulaciones humanas y que se utiliza para formular medidas de verosimilitud para la evaluación de candidatos. Los autores adoptan una programación dirigida por datos para buscar en el espacio de soluciones de manera eficiente. Además introducen la técnica que ellos denominan *proposal maps* que es una forma eficiente de implementar las propuestas de inferencia. Los resultados cualitativos y cuantitativos muestran que la técnica es eficaz en la estimación de la pose 3D en una variedad de imágenes.

En el trabajo de Mori *et al.* [28] demuestran cómo utilizar la segmentación de bajo nivel para la estimación de poses de personas. Los segmentos y superpíxeles generados por el algoritmo *Normalized Cuts* se utilizan para proponer candidatos de articulaciones y torsos. Estos candidatos son verificados usando una variedad de características. Por último, la búsqueda de configuraciones consistentes de la pose se convierte en un problema de *Constraint Satisfaction* [56].

Uno de los *frameworks* más eficientes computacionalmente para el modelado y reconocimiento de objetos basados en partes, y que se ha convertido en un estándar para la estimación de la pose humana, es la revisión de estructuras pictóricas de Felzenszwalb *et al.* [12]. Dicho trabajo está motivado por los clásicos modelos de estructuras pictóricas introducidos por Fischler y Elschlager [57]. La idea básica es la de representar un objeto por un conjunto de partes dispuestos en una configuración deformable. La apariencia de cada parte se modela por separado, y la configuración deformable está representada por conexiones entre pares de partes. Estos modelos permiten descripciones cualitativas de la apariencia visual y son adecuados para problemas de reconocimiento genéricos. Una de las aplicaciones con más éxito de las nuevas estructuras pictóricas ha sido en la estimación de la pose humana. Con relación al artículo anterior, ellos hacen uso de modelos de estructuras pictóricas para encontrar instancias de un objeto en una imagen, así como el problema de aprender un modelo de objetos a partir de muestras de entrenamiento, presentando en ambos casos resultados eficientes. Ellos muestran métodos de aprendizaje de modelos que representan rostros y cuerpos humanos y con el uso de dichos modelos resultantes localizan las poses correspondientes en nuevas imágenes.

Una de las aplicaciones más famosas de estas estructuras pictóricas es el trabajo de Ramanan [29] donde enfoca la estimación de la pose humana usando una inferencia en un modelo probabilístico. Ramanan afirma que el éxito de muchos métodos de estimación de la pose humana se encuentran sobre todo en las características. Su principal contribución es la utilización de la inferencia visual en un proceso de análisis iterativo, donde secuencialmente se aprenden y afinan las características de una imagen en particular. El trabajo muestra resultados cuantitativos para la estimación de la pose humana en una base de datos de más de 300 imágenes y demuestra que el algoritmo propuesto es competitivo y supera al

estado del arte publicado hasta 2006, año de publicación del artículo.

Un uso de la propuesta de Ramanan es el trabajo de Ferrari *et al.* [18] donde proponen una reducción progresiva del espacio de búsqueda de las partes del cuerpo para mejorar en gran medida el éxito en la estimación de la pose humana. Esta reducción implica dos contribuciones: un detector genérico usando un modelo débil de pose para reducir sustancialmente el espacio de búsqueda en una imagen; y el empleo de GrabCut [58] en regiones espaciales devueltas por el detector genérico, así se reduce aún más el espacio de búsqueda. Por último, ellos proponen un modelo espacio-temporal integrado que abarca varios fotogramas de un vídeo para perfeccionar las estimaciones de la pose en cada uno de los fotogramas, resolviendo la inferencia mediante la técnica de *Belief Propagation*.

Una mejora del trabajo de Ferrari *et al.* es el trabajo de Eichner y Ferrari [30] donde se presenta un nuevo enfoque para la estimación de modelos de apariencia de las partes del cuerpo en una imagen. Basándose en la reducción del espacio de Ferrari *et al.* , dos observaciones motivan su enfoque: (i) algunas partes del cuerpo tienen una localización más estable (por ejemplo, en una imagen el torso se encuentra en su amplia mayoría debajo de la cara); (ii) los modelos de apariencia de las diferentes partes del cuerpo están estadísticamente relacionados. Por ejemplo, los antebrazos de una persona generalmente son del color del torso (por la ropa) o del rostro (por la piel). Sólo en raras ocasiones tienen un color totalmente diferente. Esto implica que la apariencia de ciertas partes del cuerpo pueden predecirse a partir de la apariencia de otras partes. Los experimentos muestran que su técnica mejora considerablemente los resultados cuantitativos del estado del arte existentes en el momento de la publicación del trabajo.

Una recopilación y mejora de las propuestas de [18] y [30] descritas anteriormente se puede encontrar en el trabajo de Eichner *et al.* [10]. En ella presentan con éxito su metodología de estimación, en imágenes individuales, de la pose 2D de personas que aparecen a cualquier escala, en diversas condiciones de iluminación, con cualquier tipo de ropa y con indiferencia del color de piel. Al estimar poses en una sola imagen, su propuesta también está adaptada para ser utilizada en vídeos. Además de las diversas mejoras que hacen en cada una de las etapas propuestas en [18] y de la extensión al modelo de apariencia del trabajo [30], Eichner *et al.* realizan una revisión del trabajo [59] proporcionando en su artículo un marco de trabajo completo no sólo para la estimación de la pose humana sino también para la recuperación de fotogramas en vídeos a partir de una pose 2D humana dada. En el capítulo 3 de esta tesis extendemos a imágenes estéreo la estimación de la pose humana de este trabajo ([10]).

Además de los métodos basados puramente en estructuras pictóricas, hay algunos trabajos que proponen ligeras modificaciones como el de Zuffi *et al.* [31]. En dicho trabajo se define un modelo de estructuras deformables (en inglés *Deformable Structures* o DS) y es una extensión de los modelos de estructura pictórica

donde considera que las formas de las partes del cuerpo no son rígidas (es decir, no rectangulares). Un modelo de estructura deformable está representado por un espacio deformable de baja dimensión y por potenciales binarios entre las partes que reflejan cómo varía la forma con la pose y las partes vecinas. Una ventaja clave de este modelo es que se modela mejor las fronteras de los objetos. Esto permite que sus modelos de verosimilitud de imagen sean más discriminativos que los modelos de verosimilitud típicos de las estructuras pictóricas. Esta verosimilitud es aprendida mediante imágenes de entrenamiento anotadas usando estructuras deformables denominadas *títeres*. Zuffi *et al.* trabajan finalmente en un modelo de estructura deformable aprendido a partir de proyecciones 2D de un modelo de cuerpo humano en 3D y lo usan para inferir poses humanas en imágenes usando una variación de un *belief propagation* no paramétrico.

El trabajo de Yang y Ramanan [21] define una mezcla de partes del cuerpo flexibles para estimar poses humanas en imágenes monoculares. Esta mezcla de partes del cuerpo son no orientadas para modelar la variabilidad del cuerpo. Todos los parámetros del modelo son aprendidos por máquinas de vector soporte estructuradas.

Una extensión correspondiente del método anterior es la propuesta de Cherian *et al.* [19] en el que estiman la pose 2D en secuencias vídeos en un modelo con enlaces temporales. En cada fotograma generan una colección de candidatos pose que se combinan a lo largo de la secuencia utilizando información de apariencia y temporal. En el Capítulo 4 de esta tesis extendemos este trabajo para secuencias de vídeo estéreo mediante el uso de la información de disparidad de dos maneras: para segmentar personas (es decir, la eliminación de los píxeles del fondo) y para encontrar con una pose común en las dos vistas de una imagen estéreo siguiendo así la coherencia de reducción del espacio de búsqueda del Capítulo 3. Además, en el Capítulo 4 se demuestra que el caso monocular se puede mejorar la eliminación de los píxeles del fondo en base a la información de apariencia y mediante el uso de información *a priori* respecto a los hombros dada una región espacial de detección de la persona.

Existen otras propuestas que utilizan las particularidades de varias personas en una misma imagen. Uno de los trabajos a destacar es el de Eichner y Ferrari [32] donde presentan un nuevo método que extiende las estructuras pictóricas para modelar explícitamente las interacciones entre personas para así estimar sus poses de forma conjunta. Las interacciones se modelan como oclusiones entre personas. En primer lugar, proponen un predictor de oclusión basado en la localización de las personas detectadas automáticamente en la imagen e incorporan dichas predicciones como información *a priori* de oclusión a su modelo de estructura pictórica de varias personas. Por otra parte, su modelo incluye una penalización por exclusión entre personas que previene que las partes del cuerpo de diferentes personas ocupen la misma región de la imagen. Gracias a estos elementos, su

modelo tiene una visión global de la escena, lo que resulta en una mejor estimación de la pose en fotografías de grupo, donde varias personas se colocan cerca y se ocluyen entre sí. En una evaluación completa en un conjunto de fotos de grupos de personas demuestran los beneficios de su modelo para estimar la pose en una imagen comparándolo con técnicas monoculares del estado del arte que estiman a cada persona de forma independiente.

Para solventar el inconveniente del anterior trabajo cuando no se encuentran interacciones entre personas en una imagen, los mismos autores, Eichner y Ferrari, proponen un modelo de coestimación de la pose humana [22]. En este nuevo trabajo se aborda el problema de varias personas que se encuentran en una pose común pero desconocida. La tarea de su modelo es estimar las poses de forma conjunta y producir prototipos que caracterizan dicha pose compartida. Ya que las poses de cada una de las personas deben ser similares a un prototipo, el modelo propuesto tiene menos libertad en comparación con estimadores monoculares independientes, sin embargo esto simplifica el problema. En su artículo, Eichner y Ferrari muestran dos aplicaciones a su propuesta. La primera es estimar la pose de personas que realizan la misma actividad de forma sincrónica, como el aeróbic, la animación o el baile en grupo. Demuestran que su modelo mejora la precisión de la estimación de la pose sobre otros métodos monoculares independientes. La segunda aplicación es aprender prototipos a partir de un buscador de imágenes cuando se le pregunta por el nombre de una clase de pose (por ejemplo, en Google buscando *posición de loto*). Eichner y Ferrari demuestran que su modelo estima mejor la pose que otros métodos monoculares del estado de arte y aprende prototipos que pueden utilizarse como información *a priori* en estimadores de la pose humana.

En los últimos años, el aprendizaje profundo (en inglés *deep learning*) ha sido muy popular en el campo del aprendizaje automático (en inglés *machine learning*). El aprendizaje profundo intenta modelar usando arquitecturas compuestas de transformaciones no lineales y múltiples. Uno de los usos del aprendizaje profundo para la estimación de la pose humana más recientes y con éxito es el trabajo de Tompson *et al.* [33]. Su propuesta consiste en una arquitectura híbrida compuesta por una Red Neuronal Convolucional (en inglés *Convolutional Neural Networks* o CNN) de profundidad y un campo aleatorio de Markov. Ellos exponen cómo esta arquitectura se aplica satisfactoriamente al problema de la estimación de la pose humana en imágenes monoculares. La arquitectura puede trabajar con las limitaciones estructurales del dominio, como las relaciones geométricas entre las localizaciones del cuerpo. Ellos demuestran que el entrenamiento de la unión de estos dos modelos mejora el rendimiento y les permite superar enormemente a las técnicas existentes del estado del arte.

Más recientemente, se han obtenido notables resultados con el método propuesto de Pfister *et al.* [25] para la estimación de la pose humana. Este método, que se basa en las recientes y exitosas redes neuronales convolucionales como en

el trabajo anterior, asume una imagen de misma altura y anchura donde se representa una sola persona. Si bien estas restricciones sólo permite estimar la pose en unas condiciones adecuadas, es cierto que los experimentos demuestran que su método es, a día de hoy, uno de los más precisos y exactos.

Uno de los más recientes trabajos es el de Hernández *et al.* [34] donde se propone un método de *rescoring* contextual para estimar la pose humana. Los autores plantean un conjunto de *poselets* (porciones de poses) que se incorporan en el modelo y sus detecciones se utilizan para extraer características espaciales y de puntuación relativas a otras hipótesis sobre las partes del cuerpo. Ellos definen un método para la detección automática de un subconjunto compacto de *poselets* que cubre las diferentes poses en un conjunto de imágenes de validación mientras que maximiza la precisión. Un mecanismo de puntuación lo definen como un clasificador tipo *boosting* basado en conjuntos que calcula una nueva puntuación para cada detección de articulaciones. Esta nueva puntuación se incorpora en el modelo de estructura pictórica como un potencial unario adicional. Los experimentos en dos bases de datos de referencia muestran que su propuesta supera tanto en precisión como en tiempo a otros métodos del estado del arte.

2.1.2. Aproximaciones basadas en mapas de profundidad

En cuanto a la estimación de la pose humana a partir de mapas de profundidad, Zhu y Fujimura [35] proponen un método de 2 pasos donde la primera fase es el etiquetado de las partes del cuerpo humano seguida de una segunda etapa donde se realiza una estimación más precisa de la posición de las articulaciones. En el primer paso, una serie de restricciones se extraen de las características de una imagen como la cabeza y el torso. En esta etapa se aborda el problema de asignación de etiquetas, las cuales son obtenidas a partir de las restricciones. El resto de las partes superiores del cuerpo son etiquetados también en este proceso. En el segundo paso se estima las localizaciones de las articulaciones mediante restricciones cinemáticas utilizando las correspondencias entre el mapa de profundidad y las partes del modelo humano. Por último, presentan una comparación del rendimiento de su método con el estado del arte con datos de captura de movimiento.

Otro de los trabajos de estimación de la pose humana a partir de mapas de profundidad es el trabajo de Shotton *et al.* [36], en su método proponen un modelo 3D en un problema de clasificación por píxel. Su metodología consta de tres partes principales: una representación intermedia de las partes del cuerpo, un clasificador de bosque aleatorio de decisión profunda y un *pooling* entre píxeles para generar posiciones 3D del esqueleto.

El problema de trabajar en escenarios en tiempo real y la tarea de reconstrucción usando mapas de profundidad se hace difícil ya que las estrategias de optimización global tienen un gran coste computacional. Para facilitar el seguimiento

de los movimientos del cuerpo con una sola secuencia de mapas de profundidad, Baak *et al.* [37] introducen una estrategia híbrida basada en datos que combina la optimización local de la pose con técnicas globales de recuperación de la pose. La estimación final de la pose en cada fotograma se determina a partir del *track* y una hipótesis de pose donde se combinan con un esquema de selección rápida. Su algoritmo reconstruye poses complejas de todo el cuerpo en tiempo real e impide la deriva temporal, por lo que su algoritmo lo hace conveniente para varios escenarios de interacción en tiempo real.

Otra estrategia híbrida aparece en el trabajo de Ye *et al.* [38]. Este artículo presenta un sistema novedoso para estimar la pose humana a través de un sólo mapa de profundidad. El método combina detección de la pose con refinamiento de la pose. El mapa de profundidad de entrada se combina con un conjunto de ejemplos de movimientos precapturados para generar una configuración de la pose del cuerpo y un etiquetado semántico en la nube de puntos. La estimación inicial se refina directamente mediante un ajuste de la configuración de la pose con el mapa de profundidad de entrada. Además de la nueva arquitectura del sistema, sus otras contribuciones incluyen: la modificación de una técnica de suavizado de nubes de puntos para hacer frente a los mapas de profundidad ruidosos, una alineación eficiente de nubes de puntos y un algoritmo de búsqueda de pose que son independientes de la vista. Los experimentos en bases de datos públicas muestran que su enfoque logra significativamente una mayor precisión que otros métodos del estado de arte.

Otro enfoque es el propuesto por Schwarz *et al.* [39] donde usan distancias geodésicas y flujos ópticos. En este trabajo presentan un método para estimar la pose humana a partir de datos de profundidad utilizando cámaras *Time of Flight* (ToF) o con dispositivos Kinect. Su enfoque consiste en detectar puntos anatómicos en datos 3D y ajustar su modelo de esqueleto usando restricciones cinemática inversas. En lugar de depender de características basadas en la apariencia para la detección de puntos de interés que pueden variar fuertemente con la iluminación y los cambios de pose, ellos modelan a partir de una representación gráfica de los datos de profundidad que les permiten medir distancias geodésicas entre las partes del cuerpo. Ya que estas distancias no cambian con el movimiento del cuerpo, en su trabajo son capaces de localizar los puntos anatómicos independientemente de la pose. Para diferenciar las partes del cuerpo que se ocluyen entre sí, emplean información de movimiento a partir del flujo óptico entre una secuencia de imágenes de intensidad. Para terminar, realizan una evaluación cualitativa y cuantitativa de su método de seguimiento de pose en secuencias de ToF y de Kinect que contienen movimientos de diversa complejidad.

Para aprovechar las ventajas de la consistencia temporal, Sun *et al.* [40] introducen una variable latente global asociada o a la orientación del torso o a la altura de la persona. Esto aumenta la exactitud de la predicción de la pose del cuerpo.

La inclusión de distribuciones *a priori* en un modelo de bosque de regresión condicional supera a otros métodos que asumen que las localizaciones de las parte del cuerpo son independientes.

Los autores Pons-Moll *et al.* [41] introducen la Métrica Espacial de Ganancia de Información (en inglés *Metric Space Information Gain* MSIG): un nuevo árbol de decisión diseñado para optimizar directamente la entropía de la distribución en un espacio métrico. Cuando se aplica a una modelo de superficie, visto como un espacio métrico definido por distancias geodésicas, MSIG tiene como objetivo minimizar la incertidumbre en la correspondencia imagen-a-modelo. Una implementación de MSIG escalaría cuadráticamente con el número de muestras de entrenamiento. Como esto es intratable para grandes bases de datos, ellos proponen un método para calcular MSIG en tiempo lineal. Su método es una generalización de una clasificación *proxy* objectivo y no requiere un *embedding* isométrico extrínseco del modelo de superficie en el espacio euclidiano. Sus experimentos demuestran que las correspondencias que obtienen son considerablemente más precisas que las del estado del arte utilizando un número mucho menor de imágenes de entrenamiento.

El trabajo de Jiu *et al.* [42] usa redes neuronales convolucionales con bosques aleatorio de decisión para la estimación de poses humanas a partir de mapas de profundidad. En contraste con otras propuestas similares, ellos no incluyen potenciales binarios en su función de energía para así reducir los tiempos de cálculo.

2.2. Aproximaciones estéreo

Una de las primeras propuestas para estimar la pose en imágenes estéreo es el trabajo de Guo y Qian [43]. En este trabajo se presenta un *framework* basado en mezcla bayesiana experta (en inglés *Bayesian mixture expert* o BME) para la estimación de poses humanas 3D a partir de dos cámaras de separación amplia y sin calibrar. Gracias a las dos cámaras se reducen ambigüedades a la hora de estimar la pose. El BME se entrena para realizar una regresión multimodal de estimación de la pose. El algoritmo *K-means*, que tiene en cuenta la distancia euclídea y el máximo valor de distancia para el vector ángulo de la articulación, se utiliza para el clúster inicial en el aprendizaje BME. Esto proporcionará mejores clústeres para separar las poses ambiguas. También un PCA ponderado se implementa en un *framework* de maximización de la esperanza (EM) para conocer los parámetros de la BME. Esto puede reducir la dimensión de los datos de entrenamiento de manera más eficaz en comparación con un PCA global. El sistema completo es entrenado con siluetas sintetizadas a partir de datos de captura de movimiento. Los resultados experimentales con imágenes reales y sintetizadas demuestran que su enfoque no necesita de una calibración de la cámara y que estima con eficacia las poses humanas.

En el trabajo de Yang y Lee [44] se presenta un nuevo método para la recons-

trucción de la pose de un cuerpo humano en 3D a partir de secuencias de imágenes estéreo en base a un método de aprendizaje de arriba hacia abajo (en inglés *top-down learning method*). Siendo su método ineficaz para construir un modelo estadístico utilizando todos los datos de entrenamiento, ellos dividen jerárquicamente dichos datos en varios grupos para reducir la complejidad del problema de aprendizaje. En dicha etapa, se clasifican los datos de entrenamiento en varios subclústeres con imágenes de siluetas. En la etapa de reconstrucción, el método propuesto busca jerárquicamente un clúster para la mejor correspondencia en una imagen de silueta utilizando un historial de imágenes de siluetas (en inglés *silhouette history image* o SHI). A continuación, la pose 3D del cuerpo humano se reconstruye a partir de un mapa de profundidad usando una combinación lineal del método. Gracias al uso de información de profundidad, las poses similares en las imágenes de silueta se pueden estimar como diferentes poses 3D del cuerpo humano. Los resultados experimentales demuestran que el método propuesto es preciso para la reconstrucción y estimación de poses 3D del cuerpo humano.

En el artículo de Thang *et al.* [45] se presenta una técnica para estimar la pose 3D del cuerpo humano a partir de también un conjunto de secuencia de imágenes estéreo. Ellos estiman la disparidad en imágenes estéreo para reconstruir la información 3D. Los autores modelan el cuerpo humano con un conjunto de elipsoides conectados por cadenas cinemáticas y parametrizados con ángulos rotacionales en cada articulación del cuerpo. Sus principales contribuciones son: un nuevo algoritmo basado en maximización de la esperanza (EM) con iteraciones de dos pasos, la asignación de los datos en 3D a diferentes partes del cuerpo y el refinamiento de los parámetros cinemáticos para ajustar el modelo 3D a los datos. Dicho algoritmo basado en EM se itera hasta que converge en la pose correcta. Los resultados experimentales con datos sintéticos y reales demuestran que su método es capaz de reconstruir poses humanas 3D precisas a partir de imágenes estéreo.

Un trabajo que combina tres tareas como emparejamiento estéreo, segmentación y estimación de la pose sin usar cámaras infrarrojas ni modelos de cuerpos humanos muy simplificados puede encontrarse en el artículo de Sheasby *et al.* [46]. Los autores proponen un marco de trabajo para la estimación de un esqueleto humano detallado en 3D a partir de una imagen estéreo. Dentro de este *framework* definen una función de energía que relaciona los resultados de segmentación, los resultados de estimación de la pose y el mapa de disparidad. En concreto, codifican las afirmaciones de que los píxeles de primer plano: deben estar relacionados con alguna parte del cuerpo, deben corresponder a una superficie continua en el mapa de disparidad y deben estar más cerca de la cámara que los píxeles de fondo que hay alrededor. Su función de energía tiene una clase de complejidad NP-dura, sin embargo, ellos especifican cómo optimizar de manera eficiente una relajación de la misma mediante la doble descomposición (en inglés *dual decomposition*). Finalmente, demuestran que la aplicación de su propuesta conduce a mejores resultados

con respecto al estado del arte. También introducen una extensa y desafiante base de datos que usan como *benchmark* para la evaluación de estimación de la pose 3D.

En el método de Lallemand *et al.* [47] se aborda el problema de la estimación de la pose 3D a partir de imágenes estéreo. Ellos proponen un *framework* para la estimación 3D de la pose humana que se basa en clasificadores de bosques aleatorios. La primera y principal contribución es un novedoso y robusto descriptor de formas basado en rejilla y que puede ser utilizado por cualquier clasificador. La segunda contribución es un procedimiento de clasificación de dos etapas: primero se clasifica la orientación del cuerpo mediante clústeres y luego se procede a la determinación de la pose 3D dentro del clúster de orientación calculado. Para validar el método, ellos publican una base de datos de imágenes grabadas con una cámara estéreo y sincronizado con un sistema óptico de captura de movimiento que proporciona las poses *ground-truth*.

El trabajo de Seguin *et al.* [48] describe un método para obtener la segmentación en vídeos estéreo y la pose de personas en imágenes estéreo. Seguin *et al.* abordan el problema con: una tarea de etiquetado discreto de múltiples personas, una función de coste y un método de optimización eficiente. Las contribuciones de este trabajo son de dos tipos: en primer lugar, se desarrolla un modelo de segmentación que incorpora detecciones de personas y máscaras de segmentación de articulaciones aprendidas, también se añade al modelo información sobre el color y los mapas de disparidad. El modelo también representa explícitamente el orden de profundidad de cada persona y la oclusión. En segundo lugar, se introduce una base de datos de imágenes estéreo extraídas de las películas *StreetDance 3D* y *Pina*.

2.3. Aproximaciones multicámara

Uno de los trabajos más conocidos de estimación de la pose humana mediante múltiples puntos de vista es el método de Yao *et al.* [49]. Los autores defienden que el reconocimiento de acciones y estimación de la pose pueden beneficiarse mutuamente. Su enfoque consiste en un clasificador de reconocimiento de acciones que proporciona información *a priori* a un esquema de optimización basado en partículas para estimar la pose en 3D. A partir de las poses 3D estimadas y el uso de las características basadas en la pose, se lleva a cabo el reconocimiento de la acción.

La técnica de Shen *et al.* [50] estima la pose humana dada varias cámaras. Ellos trabajan en reconstrucciones de vóxeles 3D calculados a partir de siluetas 2D de primer plano en lugar de los datos de una imagen. El ajuste de plantilla (*template fitting method*) se utiliza para predecir la localización de la cabeza y el torso en estos vóxeles 3D y, a continuación, un método de ajuste jerárquico estima

las restantes partes del cuerpo.

El método de Sigal *et al.* [51] presenta un modelo *loose-limbed* que utiliza parámetros continuos para estimar la localización y la pose de una persona. Como la discretización del espacio de parámetros 3D no resulta viable, el modelo es inferido usando una variante de un algoritmo *belief propagation*.

En el trabajo de Amin *et al.* [52] se propone un modelo de estructura pictórica para múltiples vistas para estimar la pose humana en 3D. Dos contribuciones principales se presentan en este trabajo: un modelo 2D de estructura pictórica mejorado por medio de características de color y términos espaciales, y una extensión a este modelo que combina las articulaciones de las poses estimadas de los múltiples puntos de vista para inferir finalmente una pose 3D.

En el trabajo de Burenius *et al.* [53] se define una estructura pictórica 3D para estimar la pose de una persona en 3D a partir de imágenes obtenidas de varias cámaras calibradas. Para abordar el problema de la discretización del espacio de búsqueda, establecen limitaciones de la vista, esqueleto y ángulo de las articulaciones en los diferentes puntos de vista para definir una rejilla de búsqueda discreta. Por último, la pose óptima global se calcula mediante programación dinámica. Un notable uso de esta rejilla de búsqueda discreta se realiza en el trabajo de Kazemi *et al.* [54]. Su metodología para estimar la pose 3D es la siguiente: en primer lugar, un modelo de bosque aleatorio clasifica cada píxel en cada vista como una parte del cuerpo o como fondo; luego, estas verosimilitudes de las partes del cuerpo son retroproyectadas a un volumen 3D. Finalmente, la inferencia de la pose en 3D se obtiene por un modelo basado en partes.

Capítulo 3

Modelo Pictórico Estéreo

3.1. Introducción

En este capítulo tratamos el problema de la estimación de la pose 2D humana en imágenes estéreo. En particular, nuestro objetivo es estimar la localización, orientación y escala de las partes superiores del cuerpo de las personas detectadas en imágenes estéreo a partir de vídeos estéreo que se pueden encontrar en internet. Para ello proponemos una extensión del método de Eichner *et al.* [10]. Nuestro método lo denominamos Estimación de la Pose Humana Estéreo (en inglés *Stereo Human Pose Estimation* o SHPE). Una sinopsis del objetivo general del capítulo puede verse en la Figura 3.1

La contribución de este capítulo es doble. En primer lugar, se propone una nueva técnica para detectar automáticamente y estimar la pose 2D de personas en imágenes estéreo. En segundo lugar, se crea una base de datos para el problema de la estimación de la pose humana en imágenes estéreo.

El resto de este capítulo se organiza de la siguiente manera: en la Sección 3.2 describimos los fundamentos básicos de las estructuras pictóricas y el trabajo de Eicher *et al.* [10], que usaremos como punto de partida; nuestra metodología se describe en la Sección 3.3 donde proponemos modelos estéreo para la detección, segmentación e inferencia de las personas. La sección 3.4 muestra los resultados experimentales y, por último, la discusión final se presenta en la Sección 3.5.

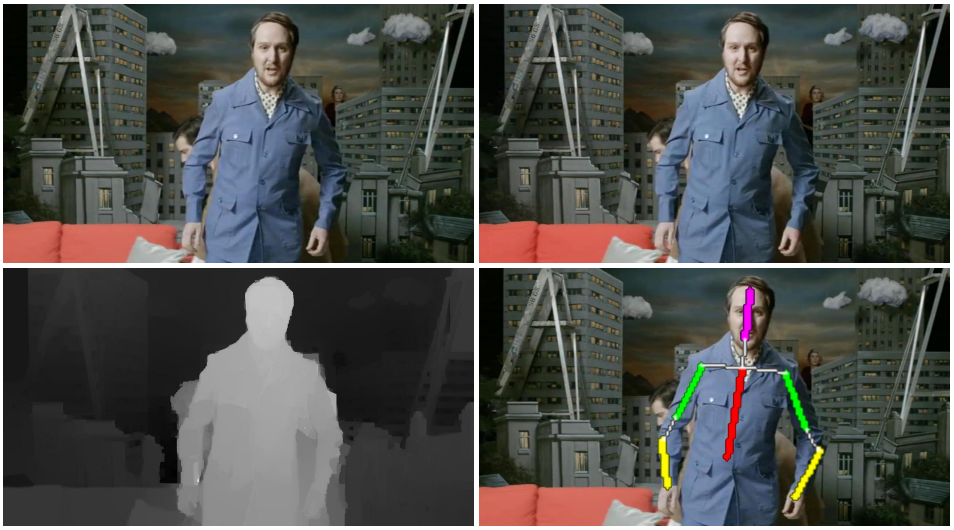


Figura 3.1: ***Objetivo del Modelo Pictórico Estéreo****. Nuestro objetivo es estimar la pose 2D de personas en vídeos estéreo. (Fila superior) Imagen estéreo de un vídeo alojado en YouTube. (Fila inferior) De izquierda a derecha, mapa de disparidad calculado a partir de la imagen estéreo y pose 2D estimada de la parte superior del cuerpo representada mediante segmentos.*

3.2. Estimación de pose humana monocular con estructuras pictóricas

Esta sección proporciona los fundamentos básicos de las estructuras pictóricas para la estimación de la pose humana.

3.2.1. Modelo de estructura pictórica

Consideremos que las partes del cuerpo de una persona están representadas por un campo aleatorio condicional [14] como lo propone [12].

Cada parte del cuerpo l_p está representado por una imagen rectangular, cuya posición está parametrizada por su localización espacial (x, y), orientación θ y escala s [12], [60]. La tupla (x, y, θ, s) constituye el espacio de estado de los nodos. La probabilidad *a posteriori* $P(L|I)$ de una configuración de las partes L dada una imagen I se define como:

$$P(L|I) \propto \exp\left(\sum_{(p,q)\in\epsilon} \Psi_{pq}(l_p, l_q) + \sum_{p} \Phi_p(I|l_p)\right). \tag{3.1}$$

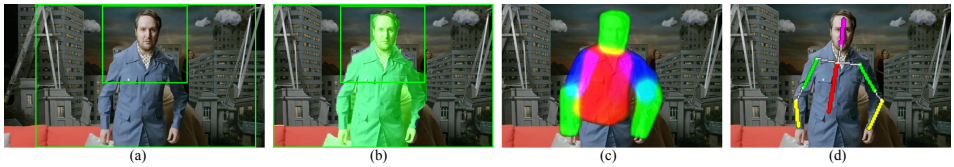


Figura 3.2: ***Propuesta de Eichner et al. [10].*** *(a) Detección de personas La salida del detector de la parte superior del cuerpo (rectángulo pequeño) se expande (rectángulo grande) para las siguientes etapas. (b) Resaltado de primer plano El resultado de la segmentación de personas elimina gran parte de innecesaria información de fondo, lo cual facilita después la búsqueda de partes del cuerpo. (c) Inferencia Los píxeles restantes son etiquetados como partes del cuerpo o como fondo. Rojo especifica el torso, azul los brazos y verde los antebrazos y cabeza. Con frecuencia, los colores se superponen; en ese caso amarillo especifica la combinación entre el antebrazo y el torso, púrpura entre el brazo y el torso, etc. (d) Ajuste de líneas. La pose 2D del cuerpo es representada por segmentos rectos (líneas) que se obtienen a partir de los resultados de (c)*

En la ecuación anterior, $\Phi_p(I|l_p)$ es el potencial asociado a la parte unaria l_p y codifica la evidencia de la imagen local de dicha parte en una posición particular (verosimilitud). Dicho potencial depende de los modelos de apariencia que describen cómo las partes se parecen. El éxito de las estructuras pictóricas para la estimación de la pose dependen fuertemente de tener buenos modelos de apariencia, lo cual limita las posiciones donde la imagen pueda contener una parte. Entre los mejores modelos nos encontramos con modelos genéricos basados en gradientes [61] y en superpíxeles [62], así como modelos específicos de persona derivados automáticamente de la imagen [29], [30].

Las restricciones cinemáticas (por ejemplo, los antebrazos deben estar unidos a los brazos) son codificadas por el potencial de pares $\Psi_{pq}(l_p, l_q)$ (es decir, una información *a priori* de la posición relativa de dos partes). Además de las restricciones cinemáticas, los potenciales de pares pueden codificar relaciones complejas como la coordinación de las partes [63] o las limitaciones de auto-oclusión [64].

La inferencia en el modelo devuelve una probabilidad máxima *a posteriori* (en inglés *maximum a posterior probability* o MAP) $L^* = \arg\max_L P(L|I)$ [12], [61] o la distribución marginal posterior para cada parte [29]. La inferencia exacta es posible cuando el modelo es un árbol [12], [18], [29], [61], sin embargo, algunos trabajos han explorado topologías más complejas [59] o mezcla de árboles [65].

3.2.2. Reducción del espacio de búsqueda

El trabajo de Eichner *et al.* [10] propone una metodología que reduce progresivamente el espacio de búsqueda de las partes del cuerpo para aumentar las posibilidades de la estimación 2D de la pose - asumiendo que el torso se limita a estar en vertical y no de perfil. Esta reducción implica: un detector de persona

genérico para reducir sustancialmente el espacio de búsqueda; y el uso de segmentación en las regiones detectadas, propuestas por el modelo anterior, para reducir aún más el espacio de búsqueda. Para terminar, su trabajo se basa en la técnica de inferencia de Ramanan [29]. Este modelo [10] se puede resumir en las etapas que se describen a continuación.

Detección y seguimiento de personas

En primer lugar, se detecta la parte superior del cuerpo humano (cabeza, hombros, etc.) en cada imagen (ver Figura 3.2(a)) utilizando un detector de ventana deslizante basado en el modelo de partes deformables de Felzenszwalb *et al.* [66]. En el caso de las secuencias de vídeo, las detecciones de la parte superior del cuerpo se agrupan en el tiempo y cada *track* resultante conecta las detecciones de una persona en cada secuencia de vídeo. Las detecciones contienen información acerca de la posición y escala de la persona en la imagen. Gracias a esta información, el conjunto de posibles localizaciones de las partes del cuerpo (x, y) se reduce y una dimensión del espacio de estados de las estructuras pictóricas, en este caso la escala, se elimina por completo. En la práctica, para cada persona detectada, el espacio de estados se limita sólo a la región de la imagen de la detección pero mediante una ligera expansión de la región espacial resultante de la detección se cubre en la medida de lo posible los brazos de una persona. Esta región espacial de la imagen se llama región espacial ampliada.

Resaltado de primer plano

En la segunda etapa de la búsqueda de las partes del cuerpo, la búsqueda se limita a la región espacial ampliada. El área de búsqueda se reduce aún más mediante la explotación de información *a priori* sobre la estructura, donde algunas áreas son muy probables que contengan partes del cuerpo mientras que otras muy poco. Esto permite inicializar la segmentación GrabCut [58] para eliminar parte del fondo (ver Figura 3.2.(b)). Por tanto, el espacio de búsqueda se limitará a las localizaciones (x, y) que se encuentran dentro del área del primer plano determinado por la segmentación GrabCut.

Estimación del modelo de apariencia

En la tercera etapa, un modelo de apariencia específico de persona [30] se aprende a partir de una sola imagen en base a dos observaciones: (i) ciertas partes del cuerpo tienen un localización bastante estable con respecto a la región espacial ampliada (esto es, cabeza y torso); y (ii) a menudo las partes del cuerpo de una persona comparten apariencias similares (por ejemplo los brazos).

Inferencia

Una pose del cuerpo se estima mediante la ejecución de la inferencia con modelos genéricos de apariencia (bordes) y modelos de apariencia específicos de persona (calculado en la tercera etapa). El área de la imagen durante la inferencia se limita a la salida del resaltado de primer plano (segunda etapa). La búsqueda explícita de las partes del cuerpo a varias escalas no es necesaria ya que la escala de la persona ha sido fijada en la primera etapa. Para cada persona detectada en la imagen, esta etapa de inferencia proporciona la distribución marginal posterior $P_i(x, y, \theta)$ para cada parte del cuerpo (ver Figura 3.2(c-d)).

3.3. Estimación de pose humana estéreo con estructuras pictóricas

El principal inconveniente del método de Eichner *et al.* [10] (resumido en la Sección 3.2.2) se puede encontrar en la etapa del resaltado de la persona. Esta etapa es crucial ya que un porcentaje de los píxeles se eliminan para su posterior procesamiento. La eliminación de partes del cuerpo, ya sea parcial o total, durante esta etapa impide que el estimador 2D de la pose sea capaz de localizar correctamente dichas partes del cuerpo. Esto sucede generalmente cuando la distribución de color de fondo es similar a algunas de las partes del cuerpo.

Como ya se ha indicado, la información adicional disponible en las secuencias estéreo se puede utilizar con el fin de superar estos problemas. Para ello, proponemos una extensión del método anterior con base a la información estéreo.

En primer lugar proponemos un detector de personas (Sección 3.3.1) que se ejecuta en ambas vistas de la imagen estéreo en todos los fotogramas del vídeo. A continuación, se aplica un algoritmo de asociación temporal para eliminar los falsos positivos. Este proceso se ejecuta de forma independiente en cada vista de las imágenes estéreo. Después adaptamos los *tracks* usando una medida basada en el grado de solapamiento de las regiones espaciales detectadas.

Luego, para cada persona detectada, se calcula la disparidad sólo en las regiones de la imagen detectadas para agilizar el proceso. Usando la información de disparidad, empleamos un método de segmentación (Sección 3.3.2) para eliminar los píxeles pertenecientes al fondo de la imagen. Por último, aplicamos nuestro modelo pictórico estéreo (Sección 3.3.3) para inferir la pose.

Como hemos dicho, sólo usamos el seguimiento temporal de Eichner *et al.* [10] para mejorar la detección de personas. Sin embargo, para nuestra etapa de segmentación e inferencia, no aplicamos restricciones temporales. Esto nos permite que nuestro modelo esté preparado para estimar la pose 2D en imágenes estéreo individuales.

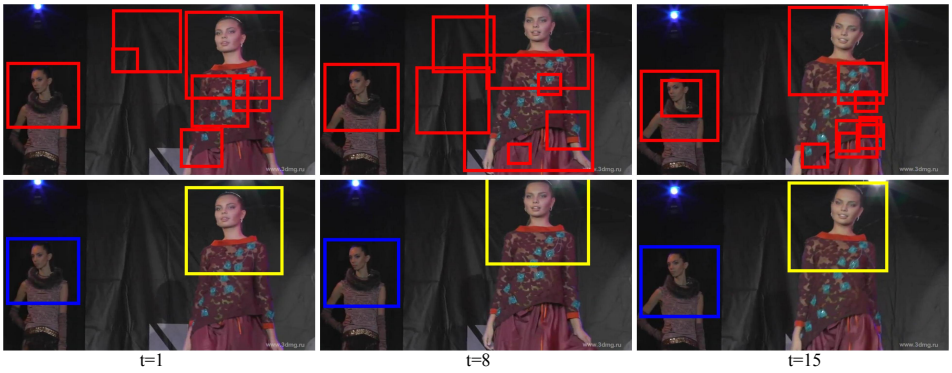


Figura 3.3: ***Detección y seguimiento de personas en Modelo Pictórico Estéreo****. Regiones espaciales devueltas por el detector de la parte superior del cuerpo (fila superior) y tracks finales (fila inferior) en fotogramas 1 (columna izquierda), 8 (columna del medio) y 15 (columna derecha) de una secuencia estéreo. Después del proceso de seguimiento, las detecciones de falsos positivos se eliminan, y un solo track se le asigna a cada persona para toda la secuencia estéreo.*

El resto de esta sección ofrece una explicación detallada de las etapas resumidas anteriormente.

3.3.1. Detección y seguimiento de personas

Al principio, al igual que hace Ferrari *et al.* en [18], empezamos por la detección de la parte superior del cuerpo humano en cada imagen para reducir el espacio de búsqueda.

Utilizamos el detector de la parte superior del cuerpo publicado por los autores de [10] en [67]. Este detector se basa en el exitoso modelo de partes deformable (en inglés *Deformable Parts Model* o DPM) de Felzenszwalb *et al.* [2]. Un DPM contiene varios filtros de histogramas de gradientes orientados [68] relacionados mediante arcos deformables.

Para empezar, se ejecuta el detector en cada vista de la imagen estéreo de forma independiente. Con el fin de eliminar los falsos positivos, se realiza un proceso de seguimiento-por-detección, como en [10], de forma independiente en cada vista de la imagen estéreo donde se genera un *track* (agrupación de detecciones de una misma persona a lo largo del tiempo) para cada persona. A los *tracks* resultantes se les da una puntuación en función de su longitud y su puntuación en las detecciones. Los *tracks* con poca puntuación se descartan para las etapas posteriores. Por último, los posibles huecos en los diferentes *tracks* (debido a fallos en la detección, por ejemplo bajo contraste o punto de vista de perfil) se completan por interpolación. La fila superior de la Figura 3.3 muestra los resultados del detector de [10] y la fila inferior el resultado final con los *tracks* calculados en una secuencia

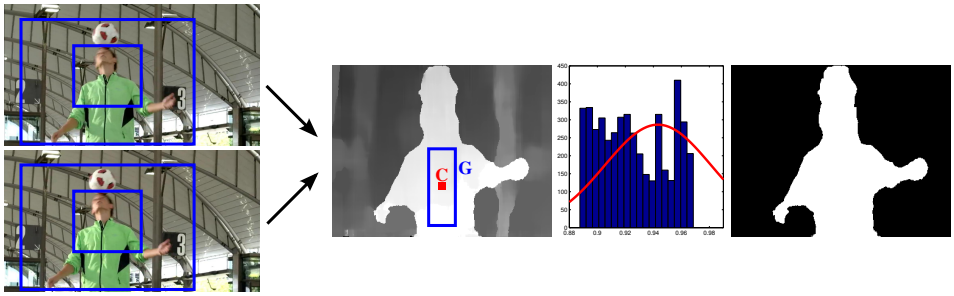


Figura 3.4: ***Detección de personas estéreo y Stereo Foreground Highlighting****. Detección estéreo de la parte superior del cuerpo:* ***(a)*** *Las dos regiones espaciales expandidas (una por cada vista) se promedian, por tanto, las regiones espaciales resultantes para ambas vistas quedan exactamente igual. Stereo Foreground Highlighting:* ***(b)*** *En primer lugar, se calcula el mapa de disparidad. Los píxeles más claros indican objetos que están más cerca de la cámara. A continuación, establecemos una región rectangular G en el torso que se utiliza como información a priori para la segmentación. El punto C es la semilla seleccionada para inicializar el algoritmo de crecimiento de regiones.* ***(c)*** *Asumimos que los valores de disparidad siguen una distribución normal, el parámetro μ se estima a partir de la región G.* ***(d)*** *Por último, la distribución aprendida previamente se utiliza para calcular la máscara binaria a partir del mapa de disparidad usando crecimiento de regiones.*

de imágenes estéreo. Obsérvese cómo los falsos positivos se eliminan después de aplicar el algoritmo.

Usando los *tracks*, anteriormente calculados, más fiables de forma independiente para cada vista, es necesario corresponderlos de una vista con otra para que coincida con la persona detectada a lo largo de la secuencia estéreo. Para ello, dado un instante de tiempo, se calcula la intersección-sobre-unión (en inglés *intersection-over-union* o IOU) [69] de todas las regiones espaciales de una vista con otra en la imagen estéreo. A continuación, se hace coincidir los *tracks* cuya suma IOU es máxima. En caso de que las cámaras diverjan mucho, se debe utilizar un procedimiento más sofisticado, como emparejamiento de histogramas de color. Sin embargo, nuestra correspondencia basada en IOU funciona ya que la mayoría de las cámaras estéreo comerciales de gama baja-media utilizadas para la grabación de vídeos tiene una separación corta entre dichas cámaras.

3.3.2. Resaltado de primer plano estéreo

Según demuestra [18], la información de localización y escala proporcionada por una detección de la parte superior del cuerpo restringe en gran medida el espacio de posibles partes del cuerpo. Con el fin de reducir aún más la búsqueda de la inferencia, el algoritmo de resaltado/segmentación de primer plano GrabCut [58]

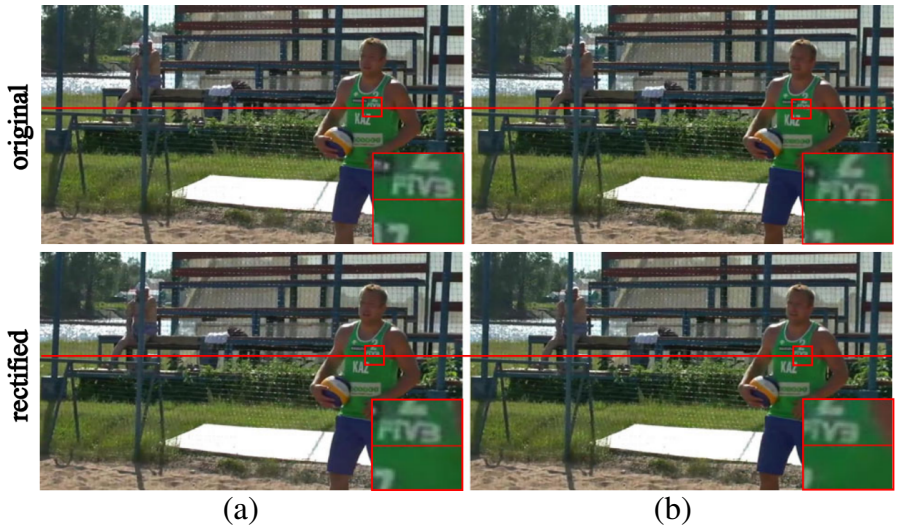


Figura 3.5: ***Rectificación en imágenes estéreo****. Las imágenes de la fila superior son la vista izquierda (a) y derecha (b) de una imagen estéreo de una de las secuencias de vídeo testadas, y las imágenes en la fila inferior muestran el resultado de aplicar nuestra rectificación. Fíjese en el zoom aplicado en las regiones cuadradas rojas. Puede verse que la línea que pasa por debajo del número* 3 *en la vista izquierda se cruza en la vista derecha. Esa desalineación vertical provoca imprecisiones en el algoritmo de block-matching estéreo durante el cálculo del mapa de disparidad. Sin embargo, los errores de alineación se reducen en gran medida, ver fila inferior, después de la rectificación.*

se amplía para extraer partes del cuerpo del fondo [10]. Sin embargo, segmentar todos los posibles brazos sigue siendo un problema difícil. Por lo tanto, proponemos una nueva estrategia para extraer a la persona a partir de una imagen, lo que ayuda, por ejemplo, a resolver el problema de la segmentación del brazo. La mayoría de los algoritmos de segmentación de imágenes se basan en un modelos de mezcla de Gaussianas en una imagen con dos clases (primer plano y fondo). Este tipo de modelos tienen una segmentación eficaz siempre que el histograma de la imagen se aproxime a una mezcla de Gaussianas y que los parámetros de dicho modelo puedan ser estimados con precisión.

Nuestra propuesta, denominada Resaltado de primer plano estéreo (en inglés *Stereo Foreground Highlighting* o SFH), explota la información estéreo para separar los píxeles pertenecientes a una persona con respecto al fondo en una imagen estéreo.

Ya que tratamos con imágenes estéreo obtenidas a través de internet, donde la información sobre la calibración de la cámara no está disponible, no podemos garantizar que las vistas de cada imagen estéreo estén debidamente rectificadas.

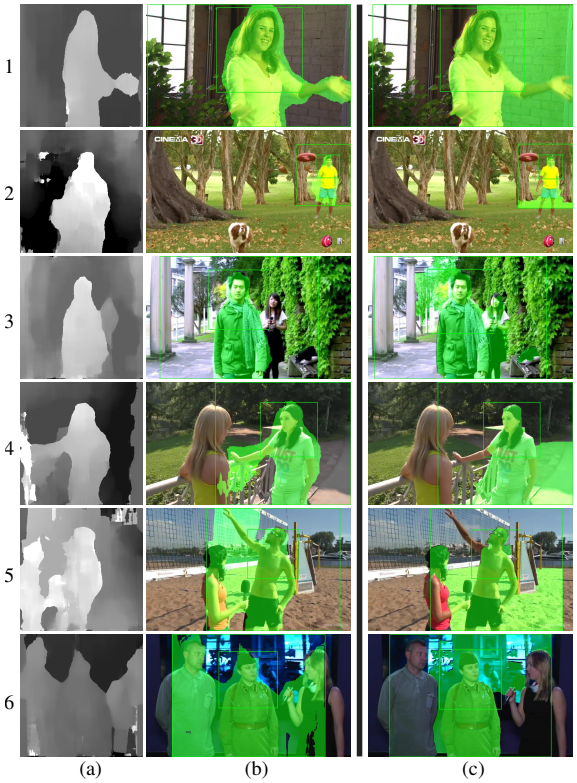
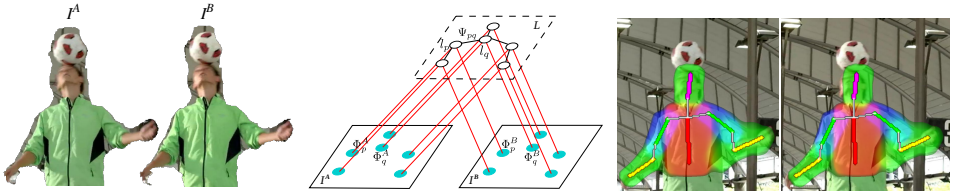


Figura 3.6: ***Resultados cualitativos de eliminación de fondo: Stereo Foreground Highlighting (SFH) vs. GrabCut****. De izquierda a derecha:* ***(a)*** *mapa de disparidad estimado para la persona objetivo;* ***(b)*** *superposición de la máscara de primer plano propuesto por SFH;* ***(c)*** *superposición de la máscara de primer plano propuesto por GrabCut tal como se utiliza en [10]. Las regiones espaciales interiores verdes en (b) y (c) representan la detección de la parte superior del cuerpo, mientras que las regiones espaciales exteriores verdes representan la región espacial ampliada resultado de aplicar nuestra etapa de detección de personas (ver Sección 3.3.1). Nótese las diferentes situaciones en la que SFH puede manejar de manera satisfactoria: los brazos en un plano diferente al del torso (por ejemplo, señalando a la cámara en la fila 1); los brazos por encima de la cabeza (fila 5); varias personas en diferentes planos de profundidad (fila 3), etc. En general, SFH elimina más píxeles de fondo que GrabCut pero manteniendo más el primer plano (por ejemplo, en la fila 4, un antebrazo no se estima como primer plano en GrabCut). Por el contrario, la fila inferior muestra un ejemplo en el que tres personas están en el mismo plano de profundidad y todas ellas se han incluido en la máscara de primer plano para segmentar sólo la persona central, mientras que GrabCut mantiene las otras personas como fondo.*

Para solucionar esto, las imágenes se rectifican utilizando el método propuesto en [70], calculando en nuestro caso la matriz fundamental (*Fundamental Matrix*) a



(a) (b) (c)

Figura 3.7: ***Inferencia en Modelo Pictórico Estéreo.*** ***(a)*** *La persona segmentada (I^A, I^B), devuelta por SFH, se utiliza para calcular $P(L|I)$.* ***(b)*** *En la estructura pictórica estéreo (SPS), cada nodo representa una parte del cuerpo para cada vista (cabeza, torso, izquierda/derecha antebrazos/brazos). El árbol incluye aristas para cada dos partes del cuerpo, parametrizados por la localización (x, y) y orientación θ, los cuales están conectados por información cinemática a priori Ψ del cuerpo humano. Cada nodo oculto (círculo vacío) se relaciona con dos nodos observados (uno por vista), representados por círculos azules los cuales se asocian a los potenciales unarios Φ (es decir, evidencias en la imagen).* ***(c)*** *Configuración de las partes del cuerpo L dadas por nuestra estructura pictórica estéreo (SPS). Obsérvese que la configuración L es la misma tanto para I^A como I^B, excepto por el desplazamiento en el eje horizontal dado por la disparidad.*

partir de un conjunto de puntos coincidentes en ambas vistas utilizando descriptores SURF [71]. Hay que tener en cuenta que la estimación de la matriz fundamental se realiza sólo una vez por cada secuencia estéreo. Un ejemplo del resultado después de aplicar dicha rectificación puede verse en la Figura 3.5. A continuación, la disparidad se calcula usando [72] y sólo en la región de la imagen establecida por la región espacial ampliada (ver Figura 3.4(a)) previamente detectada a fin de acelerar el tiempo de estimación total. El mapa de disparidad D (ver Figura 3.4(b)) indica el valor del desplazamiento horizontal en cada píxel de la vista izquierda $D_{(x,y)}$ necesario para obtener el mismo píxel en la vista derecha. Por ejemplo, un píxel (x, y) de la vista izquierda corresponde a un píxel en $(x + D_{(x,y)}, y)$ en la vista derecha.

Dado que la parte superior del torso se detecta en la fase anterior, una región rectangular más pequeña en el centro de la región ampliada o expandida se selecciona como una muestra representativa de la distribución de disparidad para todo el cuerpo en la vista izquierda (ver Figura 3.4(b)). Asumiendo que la mayoría de las configuraciones del cuerpo de una persona pueden ser modeladas como una distribución normal $\mathcal{N}(\mu, \sigma)$ (ver Figura 3.4(c)), μ se calcula como la media de los valores de disparidad en dicha región del torso. Por tanto, σ se selecciona teniendo en cuenta las dimensiones medias de las personas de manera que los brazos extendidos se ajusten en la distribución. En consecuencia, el problema de la segmentación se aborda con un algoritmo de crecimiento de regiones con semillas seleccionadas en la región de torso (véase el punto C en la Figura 3.4(b)) y con el uso de la distribución normal anterior para determinar la probabilidad de la

adición de puntos de la región segmentada . Esta propuesta permite que los brazos extendidos se capten adecuadamente como parte de la persona. La Figura 3.4(d) muestra el resultado de aplicar el método propuesto para una de las imágenes estéreo en nuestra base de datos.

Al igual que [10], también añadimos al primer plano una región rectangular, que depende de la región devuelta por el detector de la parte superior del cuerpo, que cubre la cabeza y parte del torso aprovechando así la información *a priori* proporcionada por el detector de la parte superior del cuerpo.

La máscara generada corresponde con la parte superior del cuerpo de la persona en la vista izquierda I^A. Con el fin de obtener la máscara equivalente para la vista derecha I^B, la localización de los píxeles de I^A se usan en la vista derecha aplicando la disparidad calculada. Nótese que aunque nuestro algoritmo de resaltado de primer plano estéreo comparte algunas ideas con el método propuesto por Sheasby *et al.* [73], ellos necesitan un estimador de la pose humana [74] para definir dos semillas en su algoritmo de crecimiento de regiones. Lo que para ellos es una etapa en su método, para nosotros es nuestro objetivo final.

Tanto la estimación de la disparidad como el resaltado de primer plano se llevan a cabo de forma independiente para cada región ampliada o expandida. Por tanto, situaciones tales de personas cercanas unas a otras y en diferente plano, o gente de frente o de espaldas a la cámara son tratadas de manera satisfactoria, véase por ejemplo las filas 1, 4 y 5 de la Figura 3.6. Hemos incluido en la columna (c) la máscara de segmentación obtenida mediante el algoritmo de resaltado de primer plano de Eichner *et al.* [10]. Nótese cómo frecuentemente se pierde parte de los brazos, como en las filas 2, 4 y 5, en contraste con nuestra propuesta basada en la disparidad que los mantiene satisfactoriamente. Sin embargo, en el ejemplo representado en la fila 6, debido a una estimación no muy precisa de la disparidad, nuestro método elimina menos píxeles de fondo que el enfoque basado en GrabCut.

3.3.3. Modelo pictórico estéreo (SPS)

Nuestra propuesta de añadir información estéreo en modelos pictóricos está basada en el modelo que aparece en el trabajo de Eichner *et al.* [10] resumido en la Sección 3.2.2. Entre otras aportaciones, [10] extiende el modelo de Ramanan [29] con orientaciones *a priori* Υ del torso y cabeza donde asumen que tienen una orientación cercana a la vertical.

Nosotros nos beneficiamos del algoritmo de Eichner y Ferrari [30] para generar un modelo de apariencia específico de persona Φ para cada imagen.

Las restricciones cinemáticas Ψ son las mismas que aparecen en el trabajo de Ramanan [29]: para cada posición relativa (x, y) usamos un coste truncado, dando una probabilidad uniforme cerca de la localización de la unión (*joint*) y cero en el resto de las localizaciones, y para la orientación relativa θ usamos un histograma

de orientaciones aprendido a partir de la base de datos de entrenamiento [29].

Sea I^A e I^B las vistas segmentadas después de aplicar nuestro método de Resaltado de primer plano estéreo (ver Sección 3.3.2) en una imagen estéreo (ver Figura 3.7(a)), y l_p^A, l_p^B las partes superiores del cuerpo de I^A e I^B respectivamente; dado que estamos trabajando con imágenes estéreo, se establece la siguiente relación: $l_p^B = \mathcal{D}(l_p^A, D)$ donde $\mathcal{D}(l, D)$ es una función que aplica el mapa de disparidad D a la configuración l.

Para aprovechar la información de la apariencia codificada en las imágenes estéreo, combinamos los potenciales unarios Φ_p correspondientes a la misma parte del cuerpo para cada vista a través de la función Ω. Dicha función se describe más abajo.

Por tanto, nuestra estructura pictórica estéreo para la parte superior del cuerpo (en inglés *Stereo Pictorial Model* o SPS) consiste en dos submodelos (uno por vista) relacionados por la disparidad D y la función Ω.

Cada submodelo consiste en seis partes del cuerpo denominadas cabeza, torso, brazos (izquierdo y derecho) y antebrazos (izquierdo y derecho); todos ellos conectados en una estructura de tipo árbol mediante informaciones cinemáticas *a priori* $\Psi(l_p, l_q)$.

La probabilidad de una configuración L dada por una imagen estéreo $\mathcal{I} =< I^A, I^B >$ y el mapa de disparidad D es definida por la siguiente ecuación:

$$P(L|\mathcal{I}, D) \propto \exp\Bigg\{ \sum_{(p,q)\in\epsilon} \Psi_{pq}(l_p, l_q) + \sum_p \Omega\Big(\Phi_p(I^A|l_p), \Phi_p(I^B|\mathcal{D}(l_p, D))\Big) + \Upsilon(l_{head}) + \Upsilon(l_{torso})\Bigg\}, \tag{3.2}$$

donde l_k se refiere a la configuración de la parte k en la vista A.

La selección de la función Ω conduce a instancias específicas del modelo propuesto. Uno podría pensar en la definición de Ω como, por ejemplo, la suma, el producto, la media aritmética, etc., de las verosimilitudes.

Después de llevar a cabo algunos experimentos preliminares, definimos la función Ω_{max} como el máximo de las dos verosimilitudes Φ^A y Φ^B:

$$\Omega_{max}(\Phi^A, \Phi^B) = \max\big(\Phi^A, \Phi^B\big) \tag{3.3}$$

Esta opción relaciona dos puntos de vista dando preferencia a los mayores valores de verosimilitud entre pares de puntos correspondientes.

Inferencia

El modelo pictórico estéreo (SPS) busca una configuración de las partes del cuerpo L^* que maximiza $P(L|\mathcal{I}, D)$ (ver Figura 3.7(c)):

$$L^* = \arg\max_L P(L|\mathcal{I}, D). \tag{3.4}$$

Las coordenadas (x, y) de las partes del cuerpo obtenidas en L^* se definen en el sistema de referencia de I^A. Por tanto, para obtener la localización de las partes del cuerpo en I^B se emplea la función previamente definida $\mathcal{D}(\cdot, \cdot)$.

La inferencia puede ser realizada de una manera eficiente y exacta [29], mediante el método *Belief Propagation* de suma-producto, ya que no hay bucles en el modelo gráfico (esto es, la estructura del modelo es de tipo árbol).

Detalles de implementación

Potenciales unarios Utilizamos los potenciales unitarios descritos en [10]. Los bordes de la imagen (ver Sección 3.2.2) son convolucionados con la plantilla de partes de persona-genérica publicado por el autor de [29]. Un total de 24 orientaciones discretizadas se utilizan durante la convolución para hacer frente a la rotación de las extremidades. Para los potenciales unarios basados en el color (ver Sección 3.2.2), el espacio de color CIE-Lab se utiliza para calcular histogramas de color con dimensionalidad $8 \times 16 \times 16$. Por cada parte del cuerpo l_i, tenemos una distribución de probabilidad de color c tanto para primer plano como fondo, que se utilizará como verosimilitud: $P_i(c|fg)$ y $P_i(c|bg)$.

La probabilidad *a posteriori* basada en el color (esto es, probabilidad de pertenecer a una parte i dado el píxel de color c) se calcula usando la regla de Bayes (asumiendo $P_i(fg) = P_i(bg)$):

$$P_i(fg|c) = \frac{P_i(c|fg)}{P_i(c|fg) + P_i(c|bg)}$$

Al igual que en [29], el potencial unario basado en el borde se utiliza para inicializar el modelo de color. A continuación, se añaden los dos tipos de potenciales unarios para calcular Φ_p. Remitimos al lector a [29] para más detalles.

Potenciales binarios Para el potencial binario $\Psi_{pq}(l_p, l_q)$ usamos el coste truncado como en [10], dando 0 probabilidad a configuraciones no válidas y una probabilidad uniforme a configuraciones válidas, definido por el modelo cinemático (por ejemplo la cabeza debe estar unida al torso).

Una configuración se dice que es válida tanto si la localización relativa de las extremidades como su orientación relativa están dentro de los intervalos aprendidos durante el entrenamiento. En particular, utilizamos datos previamente entrenados proporcionados por el autor de [29].

3.4. Experimentos y Resultados

Aquí se describen los experimentos que llevamos a cabo para validar la propuesta de este capítulo sobre estimación de la pose humana en vídeos estéreo.

3.4.1. Base de datos para imágenes estéreo

Nuestros experimentos se han realizado sobre nuestra base de datos para estimación de la pose: *Stereo Human Pose Estimation Dataset (SHPED)*. Más información sobre nuestra base de datos puede encontrarse en el apéndice A.1.

Con el fin de permitir resultados comparables para este capítulo sobre esta base de datos, se han definido dos particiones disjuntas en el conjunto de secuencias estéreo (esto es, 50 % cada subconjunto del total).

Estas dos particiones se han creado al azar asegurándonos de que dos secuencias estéreo extraídas de un mismo vídeo estéreo no se encuentren en la misma partición.

3.4.2. Métricas de evaluación empleadas

Nuestra técnica estima una pose 2D para cada región ampliada calculada en la base de datos SHPED. Con esta pose 2D estimada y la pose *ground-truth* anotada

(a) (b) (c)

Figura 3.8: ***Ground-truth generado para la evaluación en Stereo Foreground Highlighting (SFH)****.* ***(a)*** *Máscaras generadas automáticamente a partir de las anotaciones manuales de los segmentos de SHPED.* ***(b)*** *Salida de nuestro SFH.* ***(c)*** *Diferencias positivas de la máscara (a) menos la máscara (b). Los píxeles blancos representan píxeles ground-truth que no se han incluído en la máscara SFH. El ejemplo de la fila superior se considera una buena segmentación, mientras que el ejemplo de la fila inferior se ven que faltan píxeles en los brazos.*

en SHPED, evaluamos el rendimiento de nuestro método utilizando dos medidas.

En primer lugar, empleamos el porcentaje de partes del cuerpo correctamente estimadas (en inglés *Percentage of Correctly estimated body Parts* o PCP) que se proponen en [10].

En segundo lugar, con el fin de compendiar los valores obtenidos para los diferentes valores de τ_{PCP} utilizados para construir la curva-PCP, se calcula el área bajo la curva PCP (AUC-PCP).

Para fines de evaluación, tanto en métodos monoculares como estéreo, cada vista de la imagen estéreo se considera una instancia independiente y, por lo tanto, los dos PCP obtenidos a partir de las dos vistas no se combinan de ninguna forma.

Para una descripción más detallada de las métricas de evaluación, consultar el Apéndice B.

3.4.3. Resultados comparativos

Dado que en nuestra base de datos SHPED (ver Sección A.1) se definen dos particiones disjuntas, nuestros experimentos siguen una técnica de validación cru-

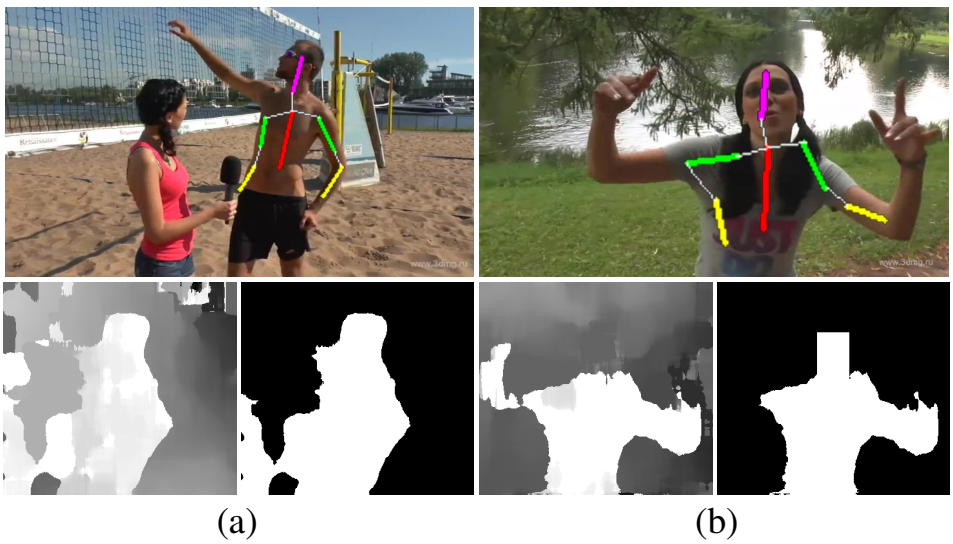


Figura 3.9: ***Casos de fallo con SHPE. (Superior)*** *Pose estimada para una sola persona.* ***(Inferior)*** *Mapa de disparidad estimado para la región de interés y máscara después de aplicar SFH.* ***(a)*** *El brazo derecho del hombre no está incluido en la máscara de primer plano debido a una estimación incorrecta de la disparidad.* ***(b)*** *Los antebrazos izquierdo y derecho y el brazo derecho no están incluidos en la máscara de primer plano después de aplicar SFH. La información a priori de cabeza + torso es claramente visible en este ejemplo en la región de la cabeza.*

Tabla 3.1: ***Siglas utilizadas en el presente capítulo.***

Acrónimo	*Nombre completo en inglés*
SPS	Stereo Pictorial Structure
SFH	Stereo Foreground Highlighting
SHPE Ω_{max}	SHPE framework using the function Ω_{max}
SHPED	Stereo Human Pose Estimation Dataset
EA [10]	Eichner *et al.* 's framework
FMP [21]	Flexible Mixtures of Parts
PCE [22]	Human Pose Co-Estimation (Direct Model)

zada de 2 iteraciones. Se calculará la media PCP y el área bajo la curva AUC-PCP en las dos particiones. Se remite al lector a la Tabla 3.1 para un resumen de los principales acrónimos utilizados en este capítulo.

Métodos de referencia

Con el fin de poner en contexto nuestros resultados, comparamos nuestro método con otros métodos monoculares del estado del arte sobre estimación de la pose. Dado que cada imagen estéreo tiene su propia anotación, tratamos cada vista de forma independiente durante la evaluación (es decir, el PCP para la vista izquierda puede diferir del PCP para la vista derecha).

Eichner *et al.* [10] (EA) Como nuestro *framework* se basa en el trabajo de Eichner *et al.* [10] (ver Sección 3.2.2), ejecutamos su algoritmo en nuestra base de datos SHPED mediante el uso de su código fuente [75] y sus parámetros por defecto. A continuación, aplicamos este método para cada vista de la imagen estéreo de forma independiente. Los resultados de este experimento se resumen en la fila 'EA' de la Tabla 3.2.

Mezcla de partes del cuerpo flexibles [21] (FMP) Yang y Ramanan proponen su método de Mezcla de partes del cuerpo flexibles (en inglés *Flexible Mixtures of Parts* o FMP) [21] para abordar el problema de estimación de la pose humana en 2D.

Debido a que FMP es considerado uno de los modelos más populares del estado del arte, ejecutamos su código fuente [76] en nuestra base de datos SHPED para fines de comparación. Utilizamos los parámetros por defecto incluídos en su software. Como se ha hecho anteriormente, aplicamos este método para cada vista de la imagen estéreo de forma independiente. Los resultados de este experimento se resumen en la fila 'FMP' de la Tabla 3.2.

Además, la fila 'FMP + BB' muestra los resultados obtenidos cuando se aplica el método FMP usando las mismas regiones devueltas por nuestra etapa de detección de personas (ver Sección 3.3.1), en lugar de buscar sobre la imagen estéreo

Tabla 3.2: ***Comparativa cuantitativa de SHPE con el estado del arte****. Aquí se muestran los resultados cuantitativos después de aplicar los diferentes métodos en SHPED. Cada entrada informa sobre los valores AUC-PCP y PCP tanto para la parte superior del cuerpo (en inglés Upper-Body Parts o UBP) como sólo los brazos (en inglés Arms). La columna AUC representa el valor AUC-PCP, Δ es la diferencia de los valores de AUC-PCP entre dicho método y EA (método de referencia), y la columna % muestra en términos de porcentaje la diferencia Δ. Con respecto a los resultados de PCP, τ_{PCP} es el umbral utilizado en la curva de PCP. Nótese que nuestro método SHPE Ω_{max} mejora claramente los resultados ofrecidos por el método de referencia EA [10], sobre todo en los brazos. Los resultados más altos están marcados en negrita.*

	AUC-PCP						*PCP* (%)			
Algoritmo	UBP			Brazos			UBPτ_{PCP}		Brazosτ_{PCP}	
	AUC	Δ	%	*AUC*	Δ	%	*0.2*	*0.5*	*0.2*	*0.5*
SHPE Ω_{max}	**0.633**	0.047	**8.0**	**0.531**	0.066	**14.2**	**50.0**	**85.6**	36.6	**79.7**
EA [10] + SFH	0.627	0.041	7.0	0.524	0.059	12.7	49.4	84.6	36.1	78.2
FMP [21]	0.599	0.013	2.2	0.509	0.044	9.5	45.7	81.4	**36.7**	72.9
FMP [21] + BB	0.589	0.003	0.5	0.505	0.040	8.6	44.2	81.3	36.2	73.0
PCE [22]	0.579	-0.007	-1.2	0.469	0.004	0.9	47.2	78.0	34.2	68.7
EA [10] (referencia)	0.586	0	0	0.465	0	0	47.3	78.6	33.1	69.4

completa. Esto permitirá una comparación mucho más directa con nuestro modelo SPS.

Coestimación de la pose humana [22] (PCE) El método Coestimación de la pose humana (en inglés Human Pose Co-Estimation o PCE) [22] intenta estimar una pose común de un grupo de personas en una imagen. Como PCE está de alguna manera relacionada con nuestro método propuesto en este capítulo (en el sentido de compartir una pose común), hemos implementado y ejecutado su método Modelo Directo, presentado en su artículo original, en nuestra base de datos SHPED. En este caso, obtenemos estimaciones de la pose comunes para cada vista de la imagen estéreo.

Utilizamos los mismos parámetros por defecto del detector de la parte superior del cuerpo y el *foreground highlighting* (resaltado de primer plano) de [10] para este método del estado del arte. Los resultados de este experimento se resumen en la fila 'PCE' de la Tabla 3.2.

Evaluación SPS

Para evaluar nuestra propuesta (ver Sección 3.3), seleccionamos los parámetros libres a fin de maximizar el AUC-PCP en el conjunto de entrenamiento – esto se repite para cada partición de la base de datos SHPED (ver Sección 3.4.1). En particular, tenemos que establecer el valor de σ para la etapa de *Stereo Foreground Highlighting* (resaltado de primer plano estéreo) (ver Sección 3.3.2). Para ello, llevamos a cabo una búsqueda por rejilla en los intervalos de $\sigma = [0.190, 0.250]$. Al igual que en el protocolo de [30], PCP y AUC-PCP se calculan sólo en las

detecciones correctas de personas (cada detección correcta cubre una pose del *ground-truth*).

En nuestro caso, las detecciones de personas cubren el 100 % del *ground-truth* de SHPED.

Nótese que utilizamos como entrada de los algoritmos monoculares nuestro conjunto de regiones espaciales de detección para todos los métodos excepto para 'FMP', tal como aparece representada en la Tabla 3.2.

La fila 'SHPE Ω_{max}' de la Tabla 3.2 muestra los resultados de nuestro modelo SPS usando la Ecuación 3.3. Nótese que mostramos los resultados AUC-PCP tanto para toda la parte superior del cuerpo (en inglés Upper-Body Parts o UBP) como para sólo los brazos y antebrazos (*arms*). Además, mostramos los resultados PCP para los valores 0.2 y 0.5 de τ_{PCP}.

Para evaluar la contribución de nuestra etapa SFH (Sección 3.3.2) en el rendimiento del sistema, en lugar de utilizar el modelo de SPS (Sección 3.3.3) para la inferencia, usamos el enfoque monocular de Eichner *et al.* [10] en la segmentación final de SFH. Los resultados de este caso se muestran en la fila 'EA+SFH' de la Tabla 3.2.

Usando nuestras anotaciones *ground-truth* de la pose, tratamos de evaluar el porcentaje de píxeles de primer plano que se pierden después de la etapa SFH (Sección 3.3.2). En la partición A perdemos alrededor de un 4.8 %, mientras que en la partición B perdemos alrededor de un 3.1 % de los píxeles de primer plano. Dos ejemplos se muestran en la Figura 3.8, en la fila superior se pierden sólo unos pocos de píxeles (ver los píxeles blancos de la máscara binaria superior de la Figura 3.8(c)); sin embargo, un porcentaje significativo de píxeles no están incluidos en el primer plano en el ejemplo representado en la fila inferior (ver los píxeles blancos de la máscara binaria inferior de la Figura 3.8(c)).

En general, las estimaciones incorrectas de la pose con SPS se deben a estimaciones inexactas de los mapas de disparidad. Dos ejemplos de dichos fallos se muestran en la Figura 3.9. En ambos casos, el modelo SPS no está en condiciones de estimar correctamente la localización de partes del cuerpo ya que han sido eliminadas previamente en la etapa de SFH (Sección 3.3.2): un brazo en el caso de la Figura 3.9(a), y dos brazos en el caso de la Figura 3.9(b).

Además de las tablas descritas anteriormente, la Figura 3.10 presenta una comparativa de curvas PCP de los métodos ejecutados en las particiones A y B (columna izquierda y derecha de la Figura 3.10, respectivamente). La fila superior de la Figura 3.10 corresponde a toda la parte superior del cuerpo (en inglés Upper-Body Parts o UBP), mientras que la fila inferior de la Figura 3.10 se corresponde con las cuatro partes del cuerpo relacionadas con los brazos (en inglés *Arms*). El valor de la métrica AUC-PCP para cada método se muestra en paréntesis en la leyenda de las gráficas.

La Figura 3.11 muestra algunos ejemplos de poses correctamente estimadas en

diversas situaciones muy complicadas, así como una estimación no tan precisa y dos fallos en situaciones difíciles (es decir, 5–b contiene una estimación muy compleja de la pose de los brazos y 5–c muestra brazos borrosos debido al movimiento).

3.4.4. Comparación con el estado del arte

Si comparamos, en términos de AUC-PCP, la fila 'EA' con la fila 'SHPE Ω_{max}' de la Tabla 3.2, podemos ver cómo nuestra metodología estéreo contribuye en gran medida al rendimiento final del sistema: un 8 %. Si nos centramos sólo en la estimación de los brazos (es decir, 4 partes del cuerpo de un total de 6), la mejora es aún mayor: un 14.2 %. En nuestra opinión, la clara mejoría mostrada por SHPE es debido a la etapa de Stereo Foreground Highlighting o resaltado de primer plano estéreo (Sección 3.3.2), donde la eliminación de píxeles del fondo es más precisa gracias a la utilización de los mapas de disparidad. De hecho, dicha etapa facilita la labor de estimación en etapas posteriores. Este hecho se refleja en la fila 'EA+SFH', donde la segmentación basada en la disparidad aumenta el rendimiento hasta en un 7 % con respecto a 'EA'.

El primer modelo que propone Eichner *et al.* [22] (ver Sección 3.4.3) para la coestimación se utiliza en nuestra comparación (fila 'PCE' de la Tabla 3.2). Sin embargo, los resultados obtenidos en este caso son aún más bajos que con los obtenidos del modelo de [10] (fila 'EA'). En nuestra opinión, aunque las regiones de detección deben estar alineadas debido a la estructura del modelo para la parte superior del cuerpo que usan, los ligeros desplazamientos existentes entre los teóricamente puntos de correspondiencia entre las dos vistas de una imagen estéreo conducen a una pobre combinación de los potenciales basados en la apariencia en su modelo de estructura pictórica durante la inferencia.

También comparamos con el exitoso modelo de Yang y Ramanan [21] (fila 'FMP' de la Tabla 3.2). Verificamos que FMP mejora sobre el la referencia de base (EA) en un 2.2 %. Sin embargo, nuestro modelo SPS mejora con respecto a FMP alrededor de un 5.7 % en toda la parte superior del cuerpo y un modesto 4.3 % en los brazos en términos de AUC-PCP. Hay que tener en cuenta que FMP utiliza un modelo articulado con subpartes a diferencia de nuestro modelo, lo que permite a FMP una estimación precisa de los brazos con respecto a otros modelos, pero con un aumento significativo en el coste computacional. Además, las partes del cuerpo pueden ser estiradas y encogidas independientemente, en contraste con nuestro modelo que tiene una escala común dada por el detector de la parte superior del cuerpo. Comparando la fila 'FMP' con 'FMP+BB', podemos decir que FMP no se beneficia de la utilización de una etapa inicial de detección de personas para limitar su búsqueda, ya que ambos resultados son muy similares. En la Figura 3.12, podemos comparar visualmente algunos de los resultados obtenidos por los modelos de SHPE y FMP. Nótese que la mayoría de los casos donde los brazos

están estirados son tratados correctamente por SHPE en contraste con FMP.

En términos de PCP (las cuatro columnas más a la derecha de la Tabla 3.2), nuestro modelo SPS alcanza un 85.6 % a 0.5 en *UBP* τ_{PCP}. Este valor es superior tanto al método de 'EA' (78.6 %) como de 'FMP' (81.4 %). Centrándonos en los brazos, FMP y SPS se comporta bastante similar a 0.2 en *Arms* τ_{PCP}.

Por último, en la Figura 3.10, podemos observar visualmente que *(i)* la curva correspondiente al modelo de SPS está, en general, por encima de los otros métodos de referencia (mejor PCP); *(ii)* FMP ofrece mejores estimaciones que EA y PCE; y, *(iii)* la estimación de la partición B es más difícil que de la partición A.

3.4.5. Tiempos de ejecución

Mostramos aquí un desglose de los tiempos de ejecución en las diferentes etapas de nuestro método propuesto dada una región espacial de la parte superior del cuerpo. La implementación se ha realizado tomando como base el código fuente del método de Eichner *et al.* [10] publicado por sus autores. La mayoría del código está escrito en MATLAB con algunas funciones *mex*. El código no se ha paralelizado ni se ha optimizado y se ha ejecutado en un PC con GNU/Linux (Ubuntu 12.04 LTS), 6 GB de RAM y una CPU a 3.4 GHz. En promedio, por una región espacial de tamaño de 160×140, la etapa Stereo Foreground Highlighting (resaltado de primer plano estéreo) tarda 2.2 segundos y la etapa de inferencia con SPS tarda 8.6 segundos.

Además, la rectificación de cada imagen estéreo (con un tamaño de imagen estéreo de 1280×720 píxeles) tarda, en promedio, 1.8 segundos. Hay que tener en cuenta que esta rectificación no se lleva a cabo para cada persona, sino para cada imagen estéreo. La estimación de los parámetros de la transformación se calcula una vez por secuencia y tarda, en promedio, 16.6 segundos.

3.5. Discusión

Este capítulo ha presentado una novedosa técnica para calcular automáticamente la pose humana 2D en imágenes estéreo extraídas de secuencias de vídeo estéreo. Además, todo el trabajo presentado en este capítulo junto con nuestra base de datos SHPED (ver Apéndice A) se ha redactado y enviado en forma de artículo a la revista *Machine Vision and Applications* y ha sido aceptado (más información sobre este artículo en el Capítulo 5).

Nuestra propuesta amplía el método monocular de Eichner *et al.* [10] de tres maneras: *(i)* una adaptación al detector de personas con un algoritmo de seguimiento para secuencias de vídeo estéreo; *(ii)* un nuevo algoritmo denominado Stereo Foreground Highlighting (resaltado de primer plano estéreo) para segmentar personas mediante el uso de mapas de disparidad; y *(iii)* un nuevo modelo de

estructura pictórica estéreo (en inglés Stereo Pictorial Structure o *SPS*) que se ejecuta sobre dos vistas en una imagen estéreo para encontrar la pose humana de la parte superior del cuerpo más probable. Con el fin de probar el método propuesto, hemos creado una base de datos denominada Stereo Human Pose Estimation Dataset con anotaciones *ground-truth* y que se ha puesto *online* a disposición de la comunidad investigadora.

Los resultados obtenidos en nuestra base de datos muestran que nuestra propuesta supera favorablemente a otros métodos del estado del arte, tales como [10] y [21]. Nuestro método combina la información de las dos vistas en una imagen estéreo para obtener la mejor aproximación de la pose humana de la parte superior del cuerpo, mejorando otros métodos monoculares que se ejecutan de forma independiente en cada vista.

Por último, debe indicarse que, aunque nuestra propuesta se ha definido para la parte superior del cuerpo, se puede extender fácilmente a cuerpos completos.

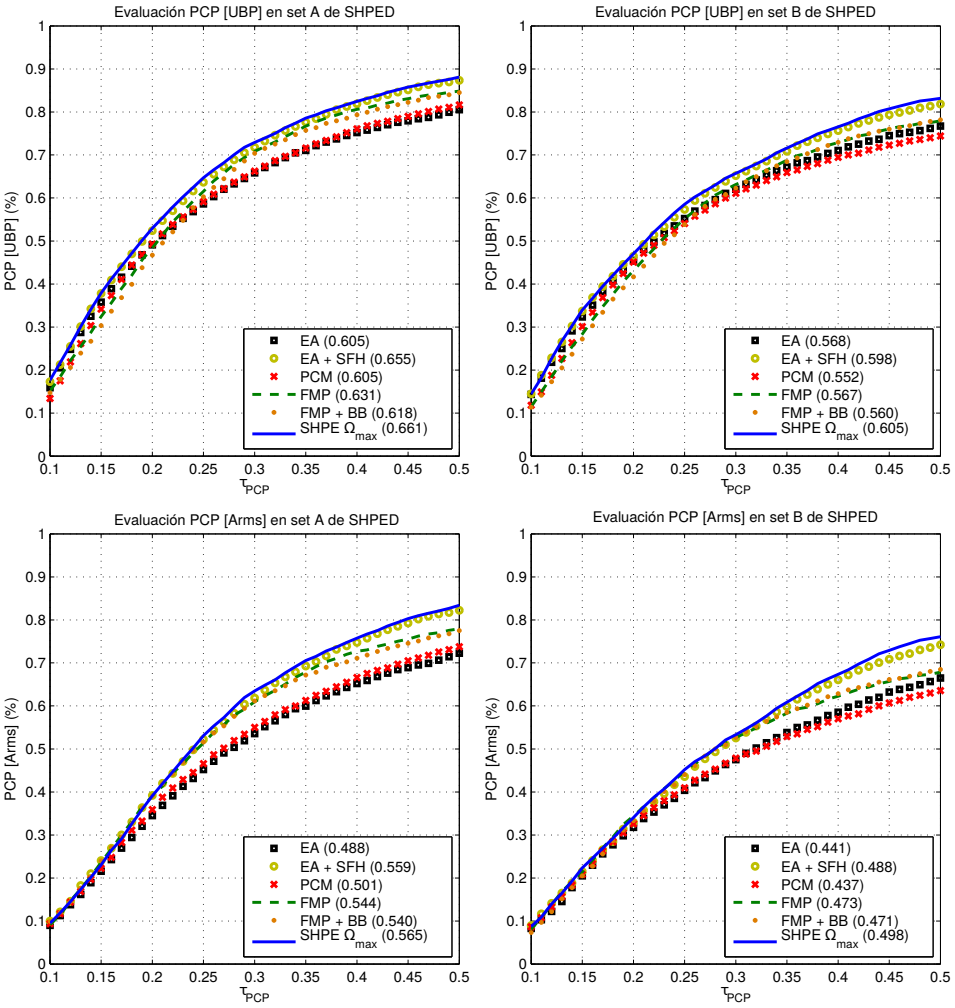


Figura 3.10: ***Comparación de curvas PCP para estimación de la pose estéreo en las particiones A y B de SHPED****. El rendimiento del framework de Eichner et al. [10] se muestra en esta figura como el método de referencia (cuadrados negros). Las curvas restantes representan tanto nuestro método propuesto en este capítulo (línea azul) como métodos de la competencia. Los valores de AUC-PCP para cada método se muestran en paréntesis. Todos los modelos se ejecutan de forma independiente tanto en la partición A* ***(columna izquierda)*** *como en la partición B* ***(columna derecha)*** *de SHPED; a su vez están clasificados en dos grupos: parte superior del cuerpo (en inglés Upper-Body Parts o UBP)* ***(fila superior)*** *y brazos (en inglés Arms* ***(fila inferior)****. En promedio, el mejor resultado se devuelve por nuestro método SHPE Ω_{max}. Esto es especialmente relevante en el caso de los brazos para $\tau_{PCP} = 0.5$.*

Figura 3.11: ***Resultados cualitativos en SHPED aplicando SHPE*** Ω_{max}. *Las filas 1 a 4 muestran ejemplos de éxito, mientras que* 5*a muestra un ejemplo casi exitoso,* 5*b y* 5*c muestran dos ejemplos de fracaso. Observe la variedad en las imágenes y la pose del brazo donde nuestro método cumple con éxito (por ejemplo,* ***1a–c, 2c, 3c, 4b****). Las imágenes son desafiantes, por ejemplo una persona puede cubrir solamente una pequeña proporción del área de la imagen* ***(2b, 4a)****, otras pueden aparecer a diferentes escalas* ***(3b)*** *o la iluminación varía en un amplio intervalo* ***(2a)****. A veces, hay un pobre contraste entre las personas y el fondo, lo que impide el uso de técnicas clásicas de sustracción de fondo* ***(3a, 4c)****. Examinando los casos de fracaso, encontramos que nuestro modelo puede confundirse a veces por una excesivo plegado de los brazos* ***(5b)*** *(tanto los antebrazos como los brazos ocupan prácticamente la misma región en la imagen) o cuando la cámara se está moviendo rápidamente, provocando un intenso desenfoque* ***(5c)****.*

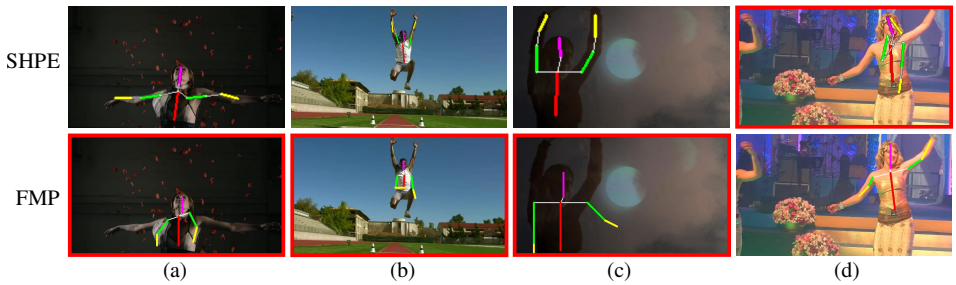


Figura 3.12: ***Comparación cualitativa entre SHPE y FMP****.* ***(Superior)*** *Poses obtenidas por nuestro método.* ***(Inferior)*** *Poses obtenidas según el modelo FMP. Las estimaciones erróneas están marcadas con un borde rojo.*

Capítulo 4

Modelo de Recombinación de Partes Estéreo

4.1. Introducción

El objetivo en este Capítulo es estimar partes del cuerpo humano en vídeos estéreo de cámaras 3D mediante el uso de información tanto de color como de disparidad. En contraposición con el anterior Capítulo 3 donde se estima la pose en imágenes estéreo sin tener en cuenta información temporal, la propuesta que se ofrece en este Capítulo se beneficia de secuencias estéreo para estimar con más precisión la pose humana.

La idea principal se muestra en la Figura 4.1 en donde primero se segmenta la persona (en cada una de las vistas) mediante el uso de la disparidad y de la información de la apariencia; después, con la ayuda de información *a priori* sobre la posible localización de los hombros, se infieren las partes del cuerpo; finalmente, se combinan las partes del cuerpo obtenidas en las diferentes vistas con la mejor configuración.

Las principales contribuciones del método propuesto en este capítulo se pueden resumir de la siguiente manera: *(i)* un conjunto de etapas para reducir el espacio de búsqueda de las partes del cuerpo en las imágenes estéreo; *(ii)* un método para combinar las partes del cuerpo en las secuencias estéreo; *(iii)* una metodología para estimar las poses de múltiples personas en imágenes estéreo; y, *(iv)* un estudio experimental exhaustivo usando tres bases de datos del actual estado del arte (*INRIA Movie 3D*, *Poses in the Wild*, *Stereo Human Pose Estimation Dataset*) para validar nuestro trabajo. Lo que pretendemos demostrar es que nuestra propuesta mejora consistentemente los resultados de otros trabajos publicados anteriormente sobre las bases de datos consideradas.

Figura 4.1: ***Objetivo de este capítulo****. Dada una secuencia de vídeo estéreo, el objetivo es localizar la posición y orientación de las partes del cuerpo humano. La disparidad e información temporal se utiliza para seleccionar y combinar las configuraciones más estables a lo largo de la secuencia.*

El resto del capítulo se organiza de la siguiente manera. Nuestra propuesta completa para la estimación partes del cuerpo en las secuencias estéreo se describe en la Sección 4.2. A continuación, la Sección 4.3 contiene los experimentos y resultados. Por último, se presenta la discusión de dichos resultados en la Sección 4.4.

4.2. Propuesta

Se describe en esta sección la metodología que seguimos para estimar poses humanas en 2D en secuencias de vídeo estéreo.

Dada una ventana de imagen que contiene una sola persona, devuelto por un detector de personas (ver Sección 4.2.2), los pasos que proponemos para la estimación de la pose en 2D son: (*i*) eliminar los píxeles del fondo mediante el uso de la disparidad e información de la apariencia (Sección 4.2.3); (*ii*) estimar la pose de forma independiente en cada parte de la imagen estéreo de cada secuencia (Sección 4.2.1); y, (*iii*) combinar las poses estimadas de forma independiente en una única pose común en cada imagen estéreo mediante la imposición de restricciones dadas por los datos estéreo (ver Sección 4.2.4).

A continuación, se comienza con un resumen de la propuesta de Cherian *et al.* [19] que estima la pose de partes del cuerpo en secuencias de vídeo monoculares. A continuación, describiremos nuestros cambios en su método para tomar ventaja de la información estéreo para mejorar la estimación de la pose.

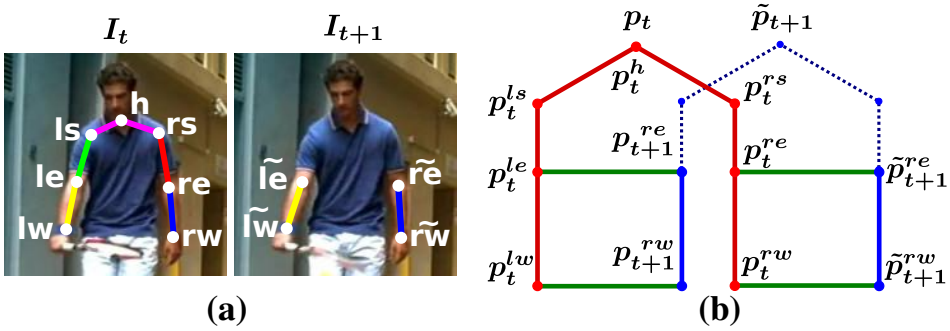


Figura 4.2: ***Modelo gráfico de Mezcla de Secuencias de Partes del Cuerpo.*** *(a) Estimación de partes del cuerpo (cabeza ([h]ead), hombros, codos y muñecas a la izquierda y derecha ([l]eft and [r]ight [s]houlders, [e]lbows and [w]rists) de I_t usando aristas temporales en I_{t+1} en codos y muñecas. (b) Modelo gráfico conectando articulaciones en dos fotogramas consecutivos (I_t, I_{t+1}).*

4.2.1. Mezcla de secuencias de partes del cuerpo

Sea $\mathcal{I} = (I_1, I_2, ..., I_T)$ una secuenca de vídeo de longitud T, donde I_i representa un fotograma del vídeo (es decir, una imagen estéreo).

El objetivo del método propuesto por [19] es estimar la pose de las partes del cuerpo (cabeza, hombros, codos y muñecas) de personas detectadas en una secuencia de vídeo (Figura 4.2.(a)).

Para este propósito, ellos define un modelo gráfico $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, donde $\mathcal{V}$ son los vértices (en este caso partes del cuerpo) y $\mathcal{E}$ son las aristas (conexiones entre pares de partes del cuerpo). Una pose p con referencia a $\mathcal{G}$ está definida como un conjunto de coordenadas 2D (x^u, y^u) que representan la posición de las partes del cuerpo en una imagen:

$p = \{p^u = (x^u, y^u) \in \mathbb{R}^2 : \forall u \in \mathcal{V}\}$

Esta fórmula conlleva la función de coste $C(I, p)$ que tiene que ser minimizado para estimar la pose del cuerpo:

$$C(I, p) = \sum_{u \in \mathcal{V}} \phi_u(I, p^u) + \sum_{(u,v) \in \mathcal{E}} \psi_{u,v}(p^u - p^v), \tag{4.1}$$

donde $\phi_u(I, p^u)$ es el término de apariencia para la parte del cuerpo u en la posición p^u y $\psi_{u,v}(p^u - p^v)$ es el coste de deformación para las partes del cuerpo (u, v). Más detalles sobre estos términos pueden encontrarse en [21].

Además de las aristas definidas entre las partes del cuerpo y para imponer una consistencia temporal entre las poses del cuerpo, Cherian *et al.* introduce una arista temporal entre los pares de partes p_t^u y p_{t+1}^u. Por tanto, la nueva función de

coste para la secuencia de vídeo complete se redefeniría como:

$$C(I_T, p_T) + \sum_{t=1}^{T-1} C(I_t, p_t) + \lambda_1 \theta(p_t, p_{t+1}, I_t, I_{t+1}), \tag{4.2}$$

donde λ_1 es un parámetro de regularización y la función θ mide la consistencia entre las poses en dos fotogramas consecutivos. En particular, θ es definido como:

$$\theta(p_t, p_{t+1}, I_t, I_{t+1}) = \sum_{u \in \mathcal{V}} ||p_{t+1}^u - p_t^u - f_t(p_t^u)||_2^2, \tag{4.3}$$

donde $f_t(p_t^u)$ corresponde al flujo óptico entre fotogramas I_t e I_{t+1} en la posición p_t^u.

Como el modelo gráfico propuesto contiene bucles y, por tanto, es un problema de inferencia intratable, ellos proponen tanto una simplificación del modelo (Figura 4.2.(b)) como un enfoque en dos etapas: (i) generar un conjunto de candidatos de poses en cada fotograma; y, (ii) descomponer todo el conjunto de candidatos de poses en extremidades para luego recomponer la pose completa combinando dichas partes del cuerpo en toda la secuencia.

4.2.2. Detección de personas

El primer paso de nuestro método es delimitar la región de imagen donde se encuentra la persona, permitiendo que nuestro método pueda utilizarse en imágenes que contengan varias personas, a diferencia de [19].

Para ello, ejecutamos el método de detector de objetos propuesto por [77] denominado *Faster R-CNN* que se basa en redes neuronales convolucionales y la proposición de regiones para hipotetizar la localización de objetos. La salida del detector es un conjunto de regiones espaciales rectangulares. En nuestro caso, cada región está ligeramente ampliada con la idea de cubrir diversas configuraciones de los brazos (ej. brazos extendidos lateralmente o brazos hacia arriba).

Las etapas posteriores de nuestro método se llevarán a cabo sólo con la información de la imagen acotada por dichas regiones ampliadas.

Asimismo, dentro de la región delimitada ejecutamos el detector de personas de [10] para localizar la parte superior del cuerpo. El resultado de esta ejecución se usará en la inicialización de la segmentación basada en la disparidad (Sección 4.2.3) y para definir posiciones a priori sobre la ubicación de las articulaciones de los hombros.

4.2.3. Segmentación de personas

Para la segmentación de personas, adoptamos el término *foreground highlighting* (resaltado de la persona) definido por [18]. Este resaltado es el proceso de

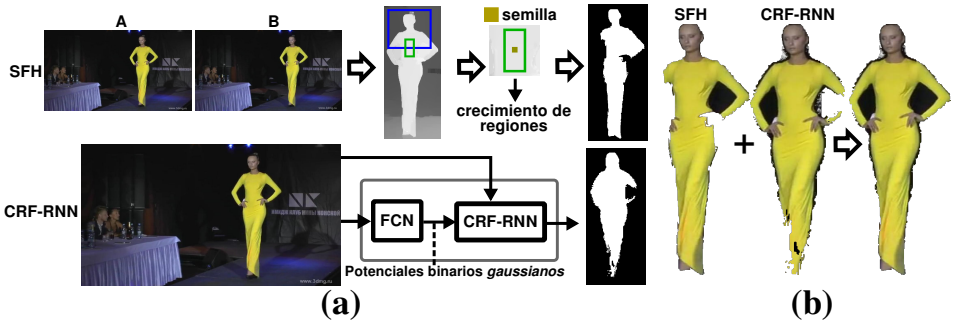


Figura 4.3: ***Segmentación en Modelo de Recombinación de Partes Estéreo.*** *(a) Se aplican dos métodos de segmentación: uno basado en la disparidad (SFH) y otro basado en la apariencia (CRF-RNN). (b) Para evitar la falta píxeles del cuerpo, se aplica un operador lógico OR a las máscaras binarias obtenidas.*

eliminación de píxeles del fondo de una imagen objetivo para limitar el espacio de búsqueda de las extremidades del cuerpo de una persona.

Proponemos aquí dos enfoques complementarios para ese propósito. El primero de ellos se basa en la disparidad estimada por píxel. La idea es aplicar un método de umbralización en el mapa de disparidad para eliminar la mayor cantidad de píxeles de fondo posible. Para tal fin, seguimos la idea del Capítulo 3 con el 'resaltado estéreo de objetos' denominado SFH (siglas en inglés de Stereo Foreground Highlighting). Recordemos que SFH es un algoritmo de crecimiento de regiones sobre el mapa de disparidad en el cual se coloca la semilla en el centro de una región en el torso. Dicha región es establecida *a priori* y está basada en el detector de parte superior del cuerpo (ver Sección 4.2.2). Se asume que los valores de disparidad del cuerpo de la persona siguen una distribución normal en donde se estiman los valores de media y varianza en la región predefinida. Un ejemplo de aplicación de este método se puede ver en la parte superior de la Figura 4.3.(a). La salida es una máscara binaria donde los píxeles del primer plano (persona) se representan de color blanco y los píxeles de fondo de color negro.

El segundo enfoque propuesto para reducir el espacio de búsqueda se basa en la información de color. El método de [78] permite asignar una etiqueta de clase (a partir de un conjunto de predefinidas) para cada píxel en la imagen objetivo. Este método, denominado CRF-RNN, se basa en una red neuronal convolucional combinada con un campo aleatorio condicional (Conditional Random Field o CRF en inglés). Para nuestro caso, estamos interesados en los píxeles etiquetados de forma automática por el método como *persona*. Se demuestra en la parte inferior de la Figura 4.3.(a) un caso típico de aplicación de este método.

Si prestamos atención a la salida de ambos métodos en la Figura 4.3, nos damos cuenta de que SFH ha eliminado el brazo derecho de la persona. Mientras

que CRF-RNN ha eliminado parte del brazo izquierdo. Sin embargo, la solución deseada sería la combinación de ambas máscaras de segmentación. Por lo tanto, combinamos las dos máscaras de segmentación mediante el operador lógico *OR*, esto es, la unión de ambos conjuntos de píxeles de primer plano. El resultado final se puede ver en la Figura 4.3.(b). Hay que tener en cuenta que esta elección es conservadora con la idea de no eliminar erróneamente píxeles pertenecientes a la persona.

4.2.4. Recombinación de partes en vídeos estéreo

Sea $\mathcal{S} = (\mathcal{I}^A, \mathcal{I}^B)$ una secuencia de vídeo estéreo, donde $\mathcal{I}^A = (I_1^A, I_2^A, ..., I_T^A)$ y $\mathcal{I}^B = (I_1^B, I_2^B, ..., I_T^B)$ corresponden las imágenes izquierda y derecha, respectivamente, de una imagen estéreo. El objetivo es encontrar una pose común del cuerpo en un determinado instante de tiempo t para cada imagen de la pareja estéreo.

Dada una parte del cuerpo $p_A^u = (x_A^u, y_A^u)$ en la imagen izquierda estará relacionada con la parte del cuerpo $p_B^u = (x_B^u, y_B^u)$ en la imagen derecha por la disparidad δ en la posición: $p_B^u = (x_A^u + \delta, y_A^u)$.

A continuación se discuten dos formas de utilización de las imágenes estéreo con el fin de mejorar la estimación de la pose a lo largo de la secuencia de vídeo estéreo.

Recombinación estéreo de extremidades

Dada una secuencia de vídeo estéreo, para cada vista de la imagen estéreo del fotograma t generamos un conjunto de K candidatos de poses $\mathcal{P}_t^A$ y $\mathcal{P}_t^B$ usando el algoritmo $n-best$ de [79], como se hace en [19]. Luego, con el fin de recombinar los candidatos de poses de cada vista en una imagen estéreo, proponemos combinar los mejores candidatos de cada vista con el fin de obtener un nuevo conjunto $\mathcal{P}_t = \left\{ \mathcal{P'}_t^A \bigcup (\mathcal{P'}_t^B + \delta) \right\}$, donde $\mathcal{P}_t$ es la unión de los mejores $K/2$ candidatos de poses para cada vista de la imagen estéreo en el fotograma t. Usamos el valor de la disparidad δ para trasladar la coordenada x de $\mathcal{P}_t^B$ y así ajustar el desplazamiento horizontal.

Una vez que los mejores candidatos de poses desde cada punto de vista han sido combinados en el nuevo conjunto, utilizamos programación dinámica para minimizar la función de coste definida en la ecuación 4.2.

Finalmente, se aplica la recombinación de partes del cuerpo de [19] y se corrige el cambio hecho en B restando los valores de disparidad añadidos previamente.

Este método lo denominamos Recombinación Estéreo de Extremidades (del inglés *Stereo Limb Recombination* o SLR) y se resume gráficamente en la Figura 4.4 (arriba).

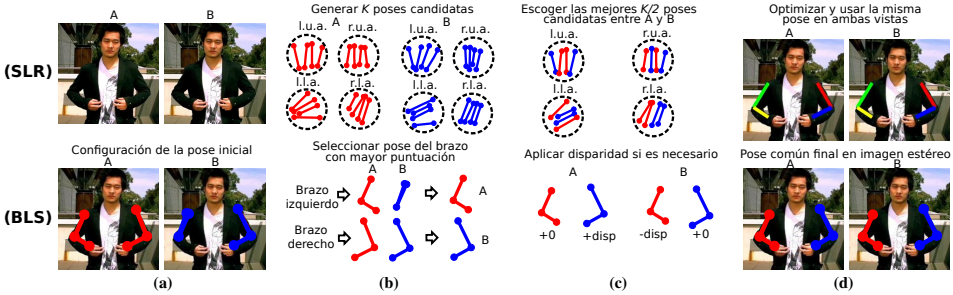


Figura 4.4: ***Mezcla de Partes del Cuerpo en Secuencias Estéreo.*** *(arriba) Recombinación Estéreo de Extremidades: los candidatos de poses de partes del cuerpo para cada vista de la imagen estéreo se combinan antes de la optimización global. (abajo) Extremidad mejor puntuada: las partes del cuerpo de cada vista de la imagen estéreo se combinan después de optimizar cada vista independientemente.*

Extremidad Mejor Puntuada

Nos centramos ahora en los brazos, donde se encuentra generalmente la mayor variabilidad.

Después de ejecutar el método monocular completo para cada vista de la imagen estéreo, una forma alternativa de aprovechar la información estéreo es seleccionar las poses candidatas con las puntuaciones más altas de cada vista.

La idea aquí es seleccionar de forma independiente la mejor configuración de las poses en cada vista y combinarlas en una sola configuración para la pareja estéreo. Para ello, en cada una de las vistas, izquierda y derecha, seleccionamos la pose de los brazos $\mathcal{P}_{arm}$ con la mejor puntuación. Por último, para cada vista trasladamos la coordenada x de $\mathcal{P}_{arm}$ con el correspondiente valor de disparidad.

Este método lo denominamos Extremidad Mejor Puntuada (en inglés *Best Limb Score* o BLS) y se resume gráficamente en la Figura 4.4 (abajo).

4.2.5. A priori sobre articulaciones

La región devuelta por el detector de la parte superior del cuerpo, en adelante BB, nos permite restringir las áreas espaciales donde los hombros deberían localizarse.

Hemos establecido dos áreas rectangulares *a priori* situadas cerca de las esquinas inferiores de BB como se ve en la Figura 4.5(a). Estas áreas tienen una relación de tamaño aproximadamente de 0.3 con respecto a BB.

Dicha información *a priori* se transfiere al proceso de estimación calculando una puntuación entre la pose inicial propuesta y la *a priori*. En particular, se calcula la relación de solapamiento v entre una región definida a partir de la posición

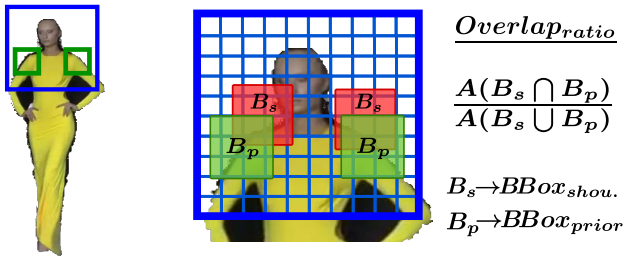


Figura 4.5: ***Aplicación de información a priori de hombros*** *(a) Dado un delimitador del detector de la parte superior del cuerpo, establecemos dos a priori de hombros (verde). (b) Superponemos nuestro delimitadores a priori B_p con las estimaciones B_s y calculamos la relacion de solapamiento (en inglés overlap ratio).*

estimada (B_s) y otra región definida a partir del *a priori* (B_p) de la siguiente manera:

$$v = \frac{area\,(B_s \cap B_p)}{area\,(B_s \cup B_p)}, \tag{4.4}$$

donde B_s y B_p corresponden a los delimitadores de la estimación de los hombros inicial y del *a priori*, respectivamente. El valor de v está en $[0, 1]$ (donde el valor 1 representa un solapamiento perfecto). Finalmente, sumamos este valor a la puntuación del candidato de pose, incrementando su probabilidad.

Un ejemplo de este solapamiento puede verse en la Figura 4.5.(b).

4.3. Experimentos y Resultados

Aquí se describen los experimentos que llevamos a cabo para validar la propuesta de este capítulo de estimación de la pose humana en vídeos estéreo.

Nuestros experimentos se han realizado sobre tres bases de datos del estado del arte para estimación de la pose: *Stereo Human Pose Estimation Dataset (SHPED)* (ver Apéndice A.1), *Poses in the Wild* (ver Apéndice A.2) e *INRIA 3D Movie Dataset* (ver Apéndice A.3).

4.3.1. Detalles de implementación

Parámetros del modelo

Usamos los términos de apariencia entrenados y los valores de deformación de [19], disponible en línea junto con el código fuente de su método [24]. Esto nos

permitirá una comparación directa con su trabajo.

Cálculo del flujo óptico

Para el cálculo del flujo óptico en secuencias de imágenes monoculares utilizamos la técnica de [80] como se propone en [19]. Este trabajo integra las correspondencias de un descriptor de articulaciones con un método variacional.

Para una estimación precisa en los mapas de disparidad de las imágenes estéreo, utilizamos el método de [11]. Hay que recordar que los mapas de disparidad se utilizan tanto para la segmentación de personas en el método SFH (Sección 4.2.3) como para ajustar el desplazamiento horizontal de las poses estimadas (Sección 4.2.4) en una imagen estéreo.

Métricas de evaluación

Utilizamos dos métricas de evaluación para comparar los diferentes métodos del estado de arte. Utilizamos dos ya que dichos métodos con los que comparamos no usan la misma métrica: error de localización en puntos clave (en inglés *Keypoint Localization Error* o KLE) (ver Apéndice B.2) y precisión media de puntos clave (en inglés *Average Precision of Keypoints o APK*) (ver Apéndice B.3).

Para SHPED y PIW, nosotros aplicamos KLE de [81]. Para cada parte del cuerpo simétrica (por ejemplo los codos, las muñecas, etc.), se evalúa el porcentaje medio (*avg*) de los lados izquierdo y derecho, y el porcentaje máximo (*max*) de dichos lados. Hay que tener en cuenta que porcentaje máximo (*max*) es la única métrica que se utiliza en [19].

Para la base de datos *INRIA 3DMovie Dataset*, utilizamos APK de [21], al igual que en [48].

4.3.2. Análisis de las diferentes etapas

Empezaremos por el estudio de la contribución en cada etapa de nuestro método sobre el rendimiento final del sistema. Para este experimento nos centramos principalmente en SHPED, ya que contiene todas las características que necesita nuestro sistema (es decir, secuencias temporales de imágenes estéreo).

En el grupo superior de filas de la Tabla 4.1 se muestra las diferentes configuraciones que evaluamos, donde 'S' indica que se ha usado la segmentación de personas de la Sección 4.2.3, 'P' indica que se ha aplicado el *a priori* de hombros de la Sección 4.2.5, y, por último, 'BLS' y 'SLR' se refieren a que se ha usado nuestros métodos de combinación de partes de la Sección 4.2.4. Hay que tener en cuenta que la fila 'MBP + BB' se refiere al método original de MBP ayudado con nuestra detección de personas (Sección 4.2.2) para hacer frente a varias personas en una única imagen (la versión del código original no dispone de esta característica).

Los resultados indican que cada una de las etapas que proponemos aporta mejoras en la MBP monocular original. La mejora más importante, con respecto a MBP en un 5 % aproximadamente, está dada por el uso del *a priori* de hombros, seguido por el uso de SLR, que aumenta la precisión de la muñeca en un 7 % aproximadamente.

Aunque la base de datos PIW no contiene imágenes estéreo, pero sí fotogramas de vídeo monocular, estudiaremos en ella la contribución de varias de nuestras etapas. Los resultados de este estudio se resumen en la Tabla 4.2. En este caso, podemos ver que nuestra etapa de segmentación 'S' (en este caso usando sólo el color, ya que no disponemos de la disparidad) trae una pequeña mejora con respecto a cada una de las partes del cuerpo.

Y, aplicando el *a priori* de hombros (fila 'S + P') la mejora es aún mayor (aproximadamente un 3 %). Por lo tanto, se puede concluir que un sistema monocular de estimación de la pose humana basado en MBP se puede mejorar mediante el uso de estas dos etapas propuestas.

Hay que tener en cuenta que si bien se trata de un hallazgo positivo, queremos recordar que el objetivo final de nuestro trabajo son las secuencias de vídeo estéreo. Por tanto, un posible nuevo estudio sobre el caso monocular queda relegado para un futuro trabajo.

Por último, las filas superiores de la Tabla 4.3 resumen los resultados de las diferentes variantes de nuestro método en las imágenes estéro de la base de datos INRIA 3D Movie. Esta base de datos no proporciona secuencias de vídeo temporales, sino imágenes estéreo aisladas. Por lo tanto, la consistencia temporal que necesita nuestro método no se puede aplicar. Sin embargo, para la completitud de nuestro estudio se muestran los resultados obtenidos mediante el uso de un modelo simplificado. Se puede ver que, como ya se observó en las otras bases de datos, el *a priori* de hombros aporta la mayor mejora con respecto a MBP (un 12 % aproximadamente en promedio). A medida que la información temporal no está disponible, BLS y SLR no ayudan a mejorar mucho los resultados.

Algunos resultados cualitativos se muestran en la Figura 4.6. Hay que tener en cuenta cómo nuestro método devuelve poses correctas en diferentes y desafiantes escenarios, incluso con múltiples personas.

4.3.3. Comparación con el estado del arte

Para propósitos de comparación, incluimos en las Tablas 4.1, 4.2 y 4.3 resultados de varios métodos del estado del arte en diferentes bases de datos. En algunos casos, hemos ejecutado el código fuente original de los autores, si estaba dicho código disponible en línea. Estos métodos se encuentran en las filas inferiores de cada tabla.

Los resultados en SHPED para varios valores de umbral de la métrica KLE se

Figura 4.6: ***Resultados cualitativos en SHPED, PIW e INRIA 3DMovie Dataset.*** *El método usado está entre paréntesis; la última columna corresponde a estimaciones erróneas. (a) SHPED ('S + P + SLR'): contiene secuencias estéreo de vídeos obtenidos en YouTube. (b) PIW ('S (Color) + P'): contiene secuencias monoculares de películas de Hollywood. (c) INRIA 3DMovie ('S + P + BLS'): contiene imágenes estéreo individuales de películas de Hollywood. Los fotogramas con borde rojo contienen estimaciones inexactas.*

resumen en la Figura 4.7, donde se representa el porcentaje de puntos significativos correctamente localizados. Además, para cada método se presenta el área bajo la curva entre paréntesis.

En el caso de los métodos de estimación de la pose monoculares, donde cada vista (derecha e izquierda) de la imagen estéreo contiene sus propias anotaciones *ground-truth*, se evalúa cada vista de forma independiente.

Una comparativa visual de los resultados obtenidos por nuestro método y por los métodos del estado del arte se presenta en la Figura 4.8. Cada fila muestra una ventana acotada de un mismo fotograma de una base de datos. La columna 'Ours' contiene nuestros resultados, especificando la combinación de etapas ejecutada según las características (es decir, monocular/estéreo, imagen/secuencia) de la base de datos.

Indicamos a continuación dónde y cómo hemos aplicado cada método comparativo.

Mezcla de secuencias de partes del cuerpo (MBP)

Usamos el código fuente proporcionado por los autores de [19], que también publicaron la base de datos PIW, para ejecutar su método MBP en las tres bases de datos consideradas. En los casos donde varias personas se muestran en una imagen, Tablas 4.1 y 4.3, hemos usado nuestro detector de personas para proporcionar a MBP la capacidad de detectar varias poses en una escena. Podemos ver que en todos los casos, nuestra propuesta mejora sobre MBP.

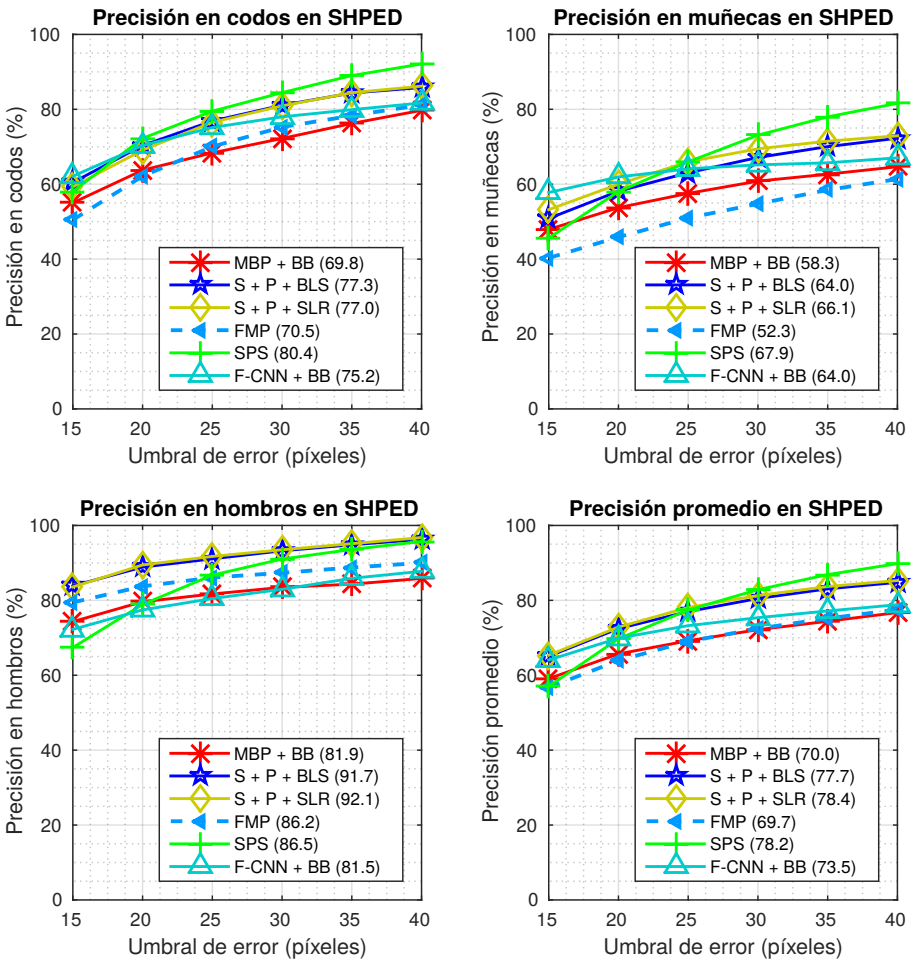


Figura 4.7: ***Comparación de curvas de precisión en la base de datos SHPED (avg.)****. Para cada curva, entre paréntesis, se muestra el área bajo la curva. Cuanto mayor sea, mejor.*

Modelo Pictórico Estéreo

El modelo pictórico estéreo (SPS) que se propone en [82], solamente se aplica en SHPED. Este método, en cuyo trabajo publicamos la base de datos SHPED, sólo se puede ejecutar en imágenes estéreo. La Tabla 4.1 muestra que nuestro método propuesto mejora en él en más de un 7 % de promedio.

Figura 4.8: ***Comparación cualitativa en SHPED, PIW e INRIA 3DMovie*** *La columna 'Ours' contiene nuestros resultados, además se indica en un recuadro blanco la configuración aplicada. Es destacable cómo nuestro método es capaz de tratar diferentes poses de brazos tanto en imágenes monoculares como estéreo.*

Mezcla de partes del cuerpo flexibles

Este método, abreviado como FMP, fue propuesto en [21] para imágenes monoculares. Podemos ver que funciona especialmente bien en la base de datos INRIA Movie 3D, superando los métodos basados en MBP. Sin embargo, tiene limitaciones claras sobre SHPED y PIW.

Redes neuronales convolucionales (ConvNets)

El método propuesto en [25], abreviado como F-CNN, se basa en redes neuronales convolucionales, y su propósito son imágenes monoculares. Este método asume sólo imágenes de entrada de misma altura que anchura y donde sólo una persona se representa. Por tanto, para experimentos sobre SHPED donde varias

Tabla 4.1: ***Comparativa de resultados cuantitativos en SHPED.*** *Comparación de precisión (en %) de localización de partes del cuerpo (15 píxeles de error de umbral). Cada entrada indica promedio y máximo (avg, max). La columna Δ corresponde al incremento de mejora o pérdida con respecto a MBP+BB.*

Método	Hombros	Codos	Muñecas	*Media*	Δ
MBP+BB	(74.3, 74.8)	(55.0, 56.5)	(47.9, 52.4)	(59.1, 61.2)	(0.0, 0.0)
S	(74.7, 75.8)	(53.7, 56.8)	(49.0, 57.1)	(59.2, 63.2)	(+0.1, +2.0)
S+P	(83.0, 84.1)	(59.5, 61.2)	(50.4, 54.5)	(64.3, 66.6)	(+5.2, +5.4)
S+P+BLS	(**83.9**, **85.0**)	(60.5, 62.0)	(50.9, 55.4)	(65.1, 67.5)	(+6.0, +6.3)
S+P+SLR	(83.6, 84.8)	(59.3, 60.9)	(53.2, 59.6)	(**65.4**, **68.4**)	(**+6.3**, **+7.2**)
FMP	(79.5, 82.0)	(50.5, 51.8)	(40.3, 41.6)	(56.7, 58.5)	(-2.4, -2.7)
SPS	(67.4, 68.4)	(58.1, 61.4)	(45.6, 48.0)	(57.0, 59.3)	(-2.1, -1.9)
F-CNN+BB	(72.2, 74.8)	(**62.1**, **64.9**)	(**57.8**, **60.8**)	(64.1, 66.8)	(+5.0, +5.6)

personas pueden estar en la misma imagen, hemos ayudado dicho método con nuestro detector de personas. Los resultados en la fila 'F-CNN + BB' de la Tabla 4.1 indican que este método es más preciso que los basados en MBP para codos y muñecas, aunque nuestro sistema completo funciona mejor en promedio total. Un comportamiento similar se observa en la base de datos PIW (Tabla 4.2). Como este método asume secuencias temporales, podemos ver en la Tabla 4.3 que obtiene resultados bajos en imágenes individuales.

Estimación de la pose combinada con segmentación

Los resultados sobre la base de datos INRIA 3DMovie son los mostrados por los mismos autores de dicha base de datos [48]. Su modelo propuesto está diseñado para imágenes estéreo donde simultáneamente realizan estimación y segmentación de personas. Como se indica en el Apéndice A , los autores no han anotado las poses en secuencias estéreo completas, sino sólo en imágenes estéreo aisladas. Este hecho se refleja en los resultados de los métodos basados en el MBP en la Tabla 4.3. Dichos métodos se basan en secuencias temporales para un rendimiento óptimo, por tanto la propuesta 'HOGcomb' de [48] muestra los mejores resultados para esta base de datos.

Tabla 4.2: ***Comparativa de resultados cuantitativos en PIW.*** *Comparación de precisión (en %) de localización de partes del cuerpo (15 píxeles de error de umbral). Cada entrada indica promedio y máximo (avg, max).*

Método	Hombros	Codos	Muñecas	*Media*
FMP	(37.4, 43.8)	(26.8, 29.7)	(19.9, 20.0)	(28.0, 31.1)
MBP	(61.2, 62.7)	(49.8, 57.0)	(42.4, 54.3)	(51.1, 58.0)
S (Color)	(65.9, 73.0)	(47.7, 58.4)	(42.7, 47.8)	(52.1, 59.7)
S (Color) + P	(**71.5**, **77.3**)	(49.5, 57.3)	(41.6, 47.8)	(54.2, 60.8)
F-CNN+BB	(68.1, 73.7)	(**55.4**, **64.0**)	(**56.3**, **58.5**)	(**59.9**, **65.4**)

Tabla 4.3: ***Comparativa de resultados cuantitativos en INRIA 3DMovie.*** *Comparación de precisión media de puntos clave (APK) (umbral $\gamma = 0.2$). Las filas que contienen la palabra 'paper' indican que los resultados han sido extraídos directamente del artículo original.*

Método	Hombros	Codos	Muñecas	*Media*
MBP+BB	0.449	0.288	0.142	0.293
S	0.500	0.310	0.134	0.315
S+P	0.706	0.393	0.144	0.415
S+P+BLS	0.758	0.383	0.144	0.428
S+P+SLR	0.758	0.389	0.139	0.429
F-CNN+BB	0.361	0.193	0.136	0.230
FMP (paper)	0.935	0.658	0.298	0.630
HOGcomb (paper)	**0.969**	**0.784**	**0.400**	**0.718**

4.4. Discusión

La metodología propuesta comienza limitando la posible ubicación de las partes del cuerpo aprovechando la información de color y disparidad de una imagen estéreo, así como la adición de información *a priori* de localización para las partes del cuerpo más estructuradas (los hombros). Finalmente, se aplica un método de recombinación de partes del cuerpo a lo largo de la secuencia estéreo para obtener la mejor configuración de pose humana. El método se ha probado sobre tres bases de datos del estado del arte: 'PIW' que contiene secuencias de vídeo monoculares, 'INRIA Movie 3D' que contiene imágenes de estereo aisladas; y, 'Stereo Human Pose Estimation Dataset' que contiene secuencias de vídeo estéreo. Este último ha sido utilizado para evaluar plenamente el impacto de nuestro trabajo, mientras que las otras dos bases de datos se han utilizado para compararlas con los métodos del estado del arte asociados a dichas bases de datos. Los resultados muestran que nuestro método obtiene mejores resultados en promedio que los comparados, estableciendo nuevos resultados en el estado del arte. Por otra parte, las etapas de reducción del espacio de búsqueda propuestas en este capítulo han demostrado ser útiles incluso en otros métodos monoculares (por ejemplo, en el trabajo Cherian *et al.* [19] - ver Tabla 4.2).

Capítulo 5

Conclusiones y Trabajo Futuro

Este capítulo presenta un resumen de la Tesis Doctoral con sus contribuciones, así como las publicaciones relacionadas con ésta y tres líneas de trabajo futuras que continuarían esta investigación.

5.1. Resumen y contribuciones de la Tesis

El objetivo principal ha sido investigar la estimación de la pose humana 2D en imágenes estéreo y demostrar cómo se pueden mejorar métodos monoculares haciendo uso de este tipo de imágenes. Para cumplirlo, primero se ha creado una base de datos de imágenes estéreo y, a continuación, se han definido dos grandes objetivos como son proporcionar dos técnicas de estimación de la pose: uno para imágenes estéreo y otro para secuencias de vídeo estéreo.

Los retos durante esta Tesis han sido varios: diferentes puntos de vista en una imagen estéreo, apariencia de las personas, ambigüedad en la poses, cambios de escala, efectos de compresión de vídeo, resolución en la imagen, oclusiones y auto-oclusiones de las partes del cuerpo y diferente iluminación y fondo en una escena.

Las principales contribuciones de esta Tesis (ver Capítulo 1.3) se resumen en:

- Un nuevo modelo de estimación de pose humana para imágenes estéreo titulado *Modelo Pictórico Estéreo*. Destaca el algoritmo de resaltado de primer plano estéreo que proporciona una mayor precisión en la estimación final de la pose que otros métodos que usan algoritmos de segmentación monoculares.

- Un nuevo modelo de estimación de la pose humana para secuencias de vídeo

estéreo nombrado *Modelo de Recombinación de Partes Estéreo*. En esta ocasión, se presentan dos submodelos para abordar el problema: *Recombinación estéreo de extremidades* y *Extremidad Mejor Puntuada*.

- Una nueva base de datos para la estimación de la pose humana denominado *Stereo Human Pose Estimation Dataset*. Esta base de datos está compuesta de 42 secuencias de vídeos estéreo extraídos de YouTube (`youtube.com`) con 15 imágenes estéreo cada secuencia.

5.2. Publicaciones relacionadas

Durante el desarrollo de esta Tesis Doctoral se han producido dos artículos que han sido enviados a revistas de ámbito científico indexadas en la publicación anual Journal Citation Reports (JCR) en la edición Science Edition.

La primera parte del trabajo de investigación, *Modelo Pictórico Estéreo* (ver Capítulo 3), fue redactado en forma de artículo. Este artículo, titulado *Stereo Pictorial Structure for 2D Articulated Human Pose Estimation*, fue enviado a la revista Machine Vision and Applications y aceptado y publicado en línea en diciembre del año 2015. En febrero de 2016 se publicó el artículo en formato de papel en dicha revista en el volumen 27, número 2.

La segunda parte del trabajo de investigación, Modelo de Recombinación de Partes Estéreo (ver Capítulo 4), fue redactado también en forma de artículo. Este segundo artículo, titulado *Mixing Body-Parts for 2D Human Pose Estimation in Stereo Videos*, ha sido enviado a una revista indexada en el JCR (*Journal Citation Reports*).

5.3. Trabajo futuro

Durante la elaboración de esta Tesis Doctoral se han logrado los objetivos propuestos y se han aportado contribuciones al estado del arte. Pudiéndose continuar el proceso de investigación, tres son los principales trabajos futuros que se plantean:

- **Estimación de la pose 3D:** la última propuesta de estimación de la pose humana en imágenes estéreo, de la que tenemos constancia, en el estado del arte, es el trabajo de Seguin *et al.* [48] y no estima la pose en 3D. Tampoco ocurre con los métodos propuestos en esta Tesis. Sin embargo, una primera aproximación se ha realizado para abordar el problema de estimación de pose 3D en imágenes estéreo. Dos resultados cualitativos obtenidos en nuestra base de datos SHPED de esta aproximación pueden verse en la Figura 5.1.

- **Ampliación de anotaciones en la base de datos SHPED:** Los elementos que se plantean para ampliar la base de datos SHPED (ver Apéndice A.1) son tres: máscaras de segmentación, imágenes negativas y anotaciones de cuerpo completo. Con máscaras de segmentación se permitiría obtener resultados cuantitativos y se vería cómo de preciso es un método de segmentación. Facilitar imágenes negativas completaría la base de datos SHPED con aquellos modelos que entrenan con este tipo de imágenes. Por último, una ampliación para SHPED sería añadir secuencias de *plano entero* (pies hacia arriba) y anotar todas las poses a cuerpo completo para todas las secuencias con dicho plano.

- **Estimación de la pose en tiempo real:** la optimización del código y, sobre todo, la paralelización mediante unidades de procesamiento gráfico (en inglés *Graphics Processor Unit* o GPU) de métodos de estimación de la pose humana resultan necesarias si se quiere estimar a tiempo real. Si se aplican estas medidas eficientemente se puede obtener, por ejemplo, con el método Modelo Pictórico Estéreo una estimación de la pose a tiempo real en una grabación en vídeo con una cámara estéreo.

Figura 5.1: ***Ejemplo cualitativo de estimación de la pose 3D****.* ***(a)*** *Pose 2D estimada.* ***(b)*** *Disparidad estimada usando [11].* ***(c)*** *Pose 3D propuesta obtenida a partir de segmentos 2D y mapa de disparidad.*

Apéndice A

Bases de datos

En esta tesis se han utilizado tres bases de datos para realizar los experimentos de estimación de la pose humana: Stereo Human Pose Estimation Dataset [82], Poses in the Wild [19] e INRIA 3DMovie Dataset [48].

La base de datos Stereo Human Pose Estimation Dataset se ha creado durante la elaboración de esta tesis. Por tanto, Stereo Human Pose Estimation Dataset forma parte de las contribuciones de esta tesis.

La complejidad de estas tres base de datos puede verse resumida en la Tabla A.1, siendo Stereo Human Pose Estimation Dataset la más completa para la realización de experimentos sobre este problema.

A.1. Stereo Human Pose Estimation Dataset

La base de datos Stereo Human Pose Estimation Dataset (abreviado SHPED) contiene 630 imágenes estéreo (esto es, 1260 imágenes en total contando las vistas derecha e izquierda) clasificadas en 42 secuencias de vídeo estéreo de 15 imágenes estéreo cada una. Las secuencias han sido extraídas de 26 vídeos estéreo, dichos vídeos han sido obtenidos del portal web YouTube (`youtube.com`) con la etiqueta

Tabla A.1: ***Complejidad de las bases de datos SHPED, PIW e INRIA 3DMovie.*** *'Estéreo': ¿la base de datos contiene imágenes estéreo? 'Multipersona': ¿la base de datos contiene anotaciones 'ground-truth' de varias personas en la misma imagen? 'Secuencias': ¿la base de datos aporta secuencias de vídeo en vez de imágenes individuales?*

Base de datos	Estéreo	Multipersona	Secuencias
Stereo Human Pose Estimation Dataset [82]	✓	✓	✓
Poses in the Wild [19]			✓
INRIA 3DMovie Dataset [48]	✓	✓	

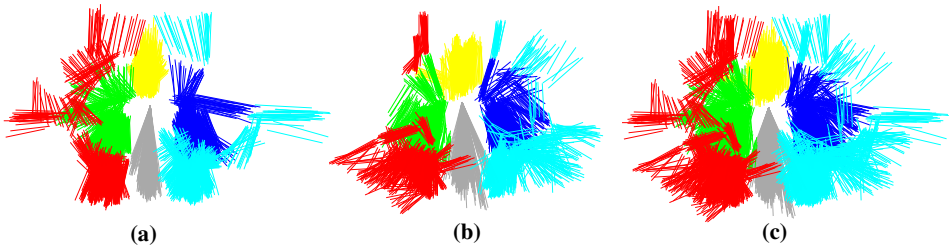


Figura A.1: ***Distribución de las poses 'ground-truth' en la base de datos SHPED****. Color de las partes superiores del cuerpo: cabeza en color amarillo; torso en color gris; brazos en colores verde y azul oscuro; y antebrazos en colores rojo y azul claro.* ***(a)*** *Partición A.* ***(b)*** *Partición B.* ***(c)*** *Base de datos completa.*

`yt3d:enable = true`.

Además, la base de datos SHPED contiene 1470 anotaciones de pose de la parte superior del cuerpo de 49 personas con las siguientes condiciones: las personas están de pie, todas las partes superiores del cuerpo están total o parcialmente visibles a lo largo de la secuencia y el punto de vista del cuerpo que no esté completamente de perfil. También incluimos transformaciones del plano proyectivas en cada secuencia para rectificar los pares estéreo y detecciones (regiones espaciales) de esas 49 personas a lo largo de toda la secuencia.

En la Figura A.2 puede verse algunas vistas de imágenes estéreo de nuestra base de datos SHPED.

Las partes superiores del cuerpo de esta base de datos están manualmente anotadas usando segmentos y puntos clave. En cuanto a segmentos anotados, las partes del cuerpo son: torso, antebrazos derecho e izquierdo y brazos derecho e izquierdo. En cuanto a puntos clave anotados, las partes del cuerpo son: hombros, codos y muñecas. En la Figura A.3 puede verse algunas anotaciones de segmentos y puntos clave. Las anotaciones están preparadas para seguir pautas de métricas de evaluación estándares como: PCP (Porcentaje de Partes del cuerpo Correctamente estimados) (ver Apéndice B.1), KLE (error de localización en puntos clave) (ver Apéndice B.2), APK (precisión media de puntos clave) (ver Apéndice B.3), etc.

Para facilitar el cálculo de parámetros libres en modelos que quieran probarse en nuestra base de datos, hemos dividido las 42 secuencias en dos particiones (21 secuencias cada partición). Con esto se puede llevar a cabo búsquedas por rejilla en unos intervalos determinados. Estas dos particiones se han dividido aleatoriamente con una condición: las secuencias de un mismo vídeo estéreo tienen que estar distribuidas equitativamente entre una partición y otra (si el número de secuencias de ese vídeo estéreo es impar, una de esas secuencias se añade a una partición completamente al azar).

Una de las maneras para determinar la calidad de una base de datos de poses

Figura A.2: ***Ejemplos de vistas de imágenes estéreo de la base de datos SHPED.*** *Aquí se muestran varias vistas de nuestra base de datos. Como se puede observar, existe una gran diversidad en la escala de personas, así como su apariencia, ropa, poses, color de piel, etc. Debido a la variedad de cámaras estéreo disponibles en el mercado, en vez de extraer imágenes estéreo de películas profesionales se han extraído de vídeos estéreo de YouTube; esto es, diferente calidad de lentes, separación entre cámaras, parámetros intrínsecos, etc. Por último, nuestra base de datos está anotada para estimaciones de varias personas en una misma vista.*

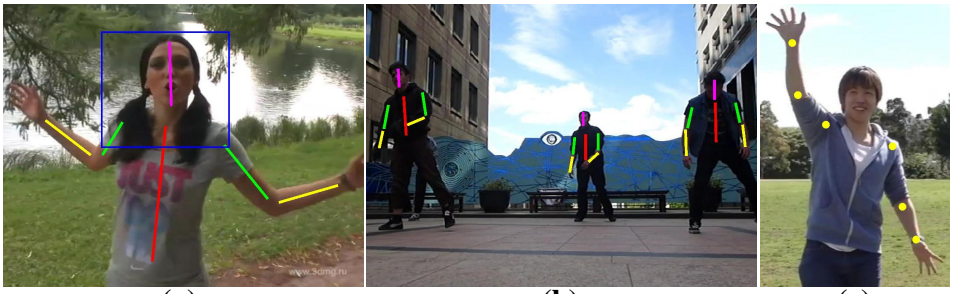


Figura A.3: ***Ejemplos de anotaciones de segmentos y puntos clave de SHPED****. (a) Segmentos anotados en una persona con su región espacial de detección. (b) Segmentos anotados en varias personas. (c) Puntos claves anotados.*

humanas es conocer la variabilidad de sus anotaciones *ground-truth*. Para ello, en la Figura A.1 muestra la variabilidad de las poses anotadas de nuestra base de datos. Se puede observar en la partición A (ver Figura A.1-a) y en la partición B (ver Figura A.1-b) cómo hay espacios sin ocupar; esto es debido a la aleatoriedad en la distribución de secuencias en las particiones. Sin embargo, la variabilidad de las anotaciones de la base de datos completa (ver Figura A.1-c) sí que demuestra una amplia diversidad de poses con menos espacios sin ocupar visibles.

La base de datos SHPED se encuentra disponible en línea [23]. La web dispone de toda la información así como la base de datos completa con todas las imágenes estéreo, transformaciones del plano proyectivas, anotaciones, identificadores de los vídeos estéreo de YouTube, así como archivos MATLAB para ejecutar demos visuales o archivos de texto como *README.md* que contiene una descripción de la base de datos y explica cómo ejecutar la demo. También ofrece las imágenes estéreo ya rectificadas (imágenes estéreo donde se le han aplicado ya las transformaciones del plano proyectivas) y en su correspondiente partición.

A.2. Poses in the Wild

La base de datos Poses in the Wild (abreviado PIW) contiene 30 secuencias de vídeo obtenidas a partir de tres películas de Hollywood: *Náufrago*, *La terminal* y *Forrest Gump*. Cada secuencia tiene aproximadamente 30 imágenes y cada imagen tiene anotados los siguientes puntos clave: cuello, hombros izquierdo y derecho, codos izquierdo y derecho, muñecas izquierda y derecha y el centro del torso.

La base de datos PIW se encuentra en línea y de modo público [24]. En dicha página web se encuentra tanto la base de datos como el código fuente de su trabajo 'Mezcla de secuencias de partes del cuerpo' [19]. Su base de datos se encuentra

en un archivo comprimido que contiene las 30 secuencias así como un archivo de texto *README* y archivos MATLAB para ejecutar demostraciones visuales.

En la Figura A.4 puede verse algunas imágenes de la base de datos PIW. Si bien tiene una diversidad en iluminación, oclusiones, ropa y poses; es cierto que sale una misma persona frecuentemente en las imágenes que es el protagonista de la película. Este inconveniente se acentúa aún más ya que el protagonista de las tres películas es el mismo. Por tanto los experimentos ejecutados en esta base de datos pueden verse afectados por la apariencia del protagonista. Si el método estima bien la apariencia del protagonista los resultados serán precisos; si no, el método no será bueno aunque estime bien otro tipo de apariencias.

A.3. INRIA 3DMovie Dataset

La base de datos INRIA 3DMovie Dataset contiene 587 poses anotadas de personas, 1158 anotaciones de regiones espaciales de personas y 686 segmentaciones de personas. Todas las imágenes estéreo han sido extraídas de las películas *Street-Dance 3D* y *Pina*. Los autores de esta base de datos eligieron estas películas ya que se grabaron con estereoscopía 3D real, a diferencia de otras películas en los que se añaden efectos 3D en postproducción y dan lugar a una estimación de la disparidad que no es la auténtica.

En el conjunto de entrenamiento están anotadas 438 poses de 232 imágenes estéreo, 520 regiones espaciales de 261 imágenes estéreo y 247 imágenes estéreo negativas sin ninguna persona en ellas. En el conjunto de test están anotadas 149 poses de 206 imágenes estéreo (la vista derecha no contiene anotaciones), 638 regiones espaciales de 193 imágenes estéreo y 686 segmentaciones de 180 imágenes estéreo.

La base de datos INRIA 3DMovie Dataset se encuentra en línea y de modo público [83]. En dicha página web se encuentra tanto la base de datos como el código fuente de su trabajo 'Pose Estimation and Segmentation of Multiple People in Stereoscopic Movies' [48]. La base de datos en su página web está organizada en tres categorías: segmentación, regiones espaciales y poses. A su vez, estas categorías tienen las siguientes subcategorías: entrenamiento y test. Para cada subcategoría se puede descargar el archivo comprimido correspondiente a la subcategoría y otro con los mapas de disparidad (calculados usando el método [11]) de sus imágenes estéreo. La base de datos también permite descargar las imágenes estéreo negativas y sus mapas de disparidad correspondiente. Por último, desde la página web se puede obtener el archivo de texto *README* y archivos MATLAB para ejecutar demonstraciones visuales.

En la Figura A.5 puede verse algunas vistas de la base de datos INRIA 3DMovie Dataset. Tiene una alta variedad en cuanto a poses, iluminación, escala, número de personas, fondo, ropa, color de la piel, etc. Sin embargo, las imágenes estéreo

están obtenidas de dos películas profesionales de Hollywood cuyas cámaras estéreo usadas durante el rodaje son de alto precio y calidad. Esto significa una calidad sobresaliente en las imágenes estéreo pero que obstaculiza la evaluación de métodos de estimación de pose en imágenes estéreo de no tanta calidad ya que han sido obtenidas de cámaras estéreo más asequibles en precio. Por último, no existe variedad con respecto a la distancia entre las lentes de las cámaras ya que sólo se obtienen las imágenes estéreo de dos películas, por tanto de dos cámaras estéreo profesionales.

Figura A.4: ***Ejemplos de imágenes de la base de datos PIW.*** *Aquí se muestran varias imágenes extraídas de la base de datos PIW. Se puede apreciar en dicha base de datos una variedad en poses así como de iluminación y fondo. Sin embargo, esta base de datos sólo está preparada para estimar una pose por imagen, impidiendo valorar la característica multipersona. Además las personas tienen una altura del torso de entre 90 y 130 píxeles aproximadamente. Esto podría ocasionar problemas a la hora de entrenar o evaluar modelos donde las personas están a diferente distancia de la cámara.*

Figura A.5: ***Ejemplos de vistas de imágenes estéreo de la base de datos Inria 3DMovie Dataset.*** *Aquí se muestran varias vistas de imágenes estéreo obtenidas de la base de datos Inria 3DMovie Dataset. En esta base de datos se puede observar la diferencia de poses, iluminación, fondo, ropa, color de piel, apariencia, etc. Si bien esta base de datos dispone de anotaciones de varias personas en una misma imagen y personas a diferente escala, ofrece sólo* 149 *anotaciones 'ground-truth' en total para la estimación de la pose humana.*

Apéndice B

Métricas de evaluación

En este apéndice se describen las métricas usadas para la estimación de la pose 2D sobre imágenes. En la Figura B.1 pueden verse ejemplos de las tres métricas usadas.

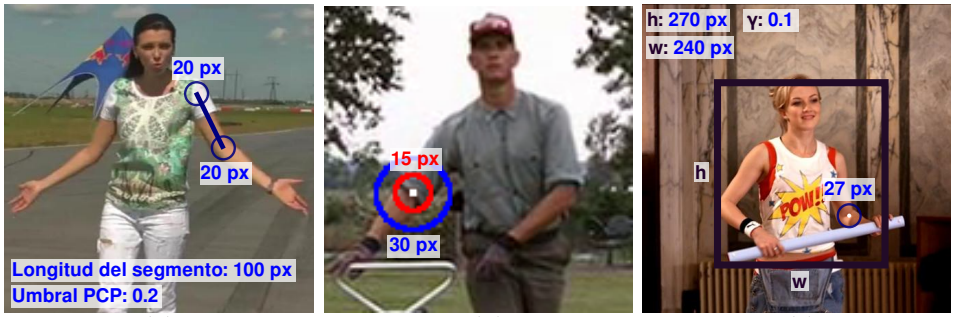


Figura B.1: ***Métricas de evaluación****. (a) PCP (un ejemplo): con una longitud del segmento ground-truth de* 100 *píxeles y un umbral de* 0.2*, una parte estimada se considera correcta si sus puntos extremos están dentro de los* 20 *píxeles de distancia de los puntos extremos del ground-truth. (b) KLE (dos ejemplos): un punto clave se considera bien estimado si está a una distancia de* 15 *píxeles, o en el otro caso* 30*, desde el punto clave ground-truth (centro en blanco). (c) APK (un ejemplo): con una región espacial ground-truth de* $(270, 240)$ *y un umbral de* 0.1*, un punto clave se considera bien estimado si está a una distancia de* 27 *píxeles, ya que* $\text{máx}(270, 240) \cdot \gamma$ *es igual a* 27*, desde el punto clave ground-truth (centro en blanco).*

B.1. PCP

El Porcentaje de Partes del cuerpo Correctamente estimadas (en inglés *Percentage of Correctly estimated body Parts* o PCP) es quizás la métrica más popular en Visión Artificial para la evaluación de estimación 2D de la pose. La medida PCP fue propuesta inicialmente por Ferrari *et al.* [18] para la evaluación de la estimación de la pose en la base de datos Buffy [84]. Según PCP, la posición de un segmento de una parte del cuerpo devuelto por un algoritmo se considera correcta si los puntos extremos de dicho segmento (proximal y distal) se encuentran dentro del $x\,\%$ de la longitud del segmento *ground-truth* desde su posición anotada (ver Figura B.1-a). Habitualmente, a la varaible x se le llama umbral PCP (en inglés *PCP threshold*) y suele tener un valor de 0.5 (50 %). En esta tesis usamos τ_{PCP} como símbolo matemático del umbral PCP.

Cuanto menor sea el umbral PCP (τ_{PCP}), más estricta es la métrica pero más precisas las partes del cuerpo estimadas que se consideran correctas.

Si representamos los valores de PCP en un intervalo de umbral PCP, normalmente entre 0.1 y 0.5, se obtiene una curva PCP. Un ejemplo de curvas PCP lo podemos ver en la Figura 3.10 del Capítulo 3.

Yang y Ramanan en su trabajo Mezcla de partes del cuerpo flexibles [21] reseña tres dificultades asociadas al uso de PCP en la práctica.

En primer lugar, el kit de herramientas de Buffy Stickmen [84] usa una definición de PCP más relajada ya que puntúa el promedio de los puntos extremos estimados en vez de puntuar los puntos extremos estimados independientemente. Por tanto, no está claro en los trabajos del estado del arte publicados qué valores PCP utilizan, si el código de evaluación del kit de herramientas o la definición original.

En segundo lugar, PCP es sensible a la cantidad de escorzo. El escorzo es la representación del segmento, en este caso, según las normas de la perspectiva. Si una parte del cuerpo tiene mucha perspectiva hacia la cámara, el segmento será más pequeño y por tanto su evaluación será más estricta.

Por último, PCP requiere que las poses candidatas y las poses *ground-truth* estén colocadas en correspondencia, pero los autores no especifican cómo se obtiene exactamente esta correspondencia.

B.2. KLE

En el error de localización en puntos clave (en inglés *Keypoint Localization Error* o KLE) se calcula el porcentaje de puntos clave que se encuentran dentro de una distancia dada al punto clave *ground-truth* [81].

Hay dos maneras de acumular el porcentaje de puntos claves estimados en una pose. La primera manera es acumular la media (*avg*) de aciertos entre partes del

cuerpo simétricas (hombros izquierdo y derecho, codos izquierdo y derecho, etc.). Por ejemplo, si un punto clave estimado del hombro izquierdo se considera correcto pero el del derecho no, la función *avg* devolvería el 50 % (0.5). La segunda manera es acumular el máximo (*max*) de aciertos entre partes del cuerpo simétricas. Por ejemplo, si punto clave estimado del codo derecho se considera correcto pero el del izquierdo no, la función *max* devolvería el 100 % (1.0).

Una de las ventajas de KLE es su fácil implementación ya que sólo se necesita calcular la distancia desde el punto clave estimado hasta el punto clave *ground-truth* y verificar si dicha distancia está por debajo de un umbral determinado.

Otra ventaja es que esta métrica es muy sencilla de entender y visualizar. El punto clave *ground-truth* es el centro de una circunferencia de radio el valor de umbral. Todo punto estimado perteneciente o interior a la circunferencia se considera correcto; en caso de que sea un punto estimado exterior a la circunferencia se considera erróneo.

KLE tiene una desventaja y es que al evaluar un conjunto de poses estimadas en una base de datos cuyas personas están a distinta escala, los resultados de dicha evaluación no estarían equilibrados. Esta sería la principal razón por la cual no se puede usar KLE directamente en nuestra base de datos SHPED (ver Apéndice A.1) sin preprocesado de datos. SHPED contiene secuencias de imágenes estéreo de personas que se encuentran a diferente distancia de la cámara lo que resulta en una altura en píxeles diferentes. Dicha diversidad de escalas hace que una secuencia, con un valor de umbral 15 píxeles, donde la persona tiene una altura en píxeles de 400 (porque está cerca de la cámara) será más difícil, por probabilidad, de estimar correctamente que otra secuencia donde la persona tiene una altura en píxeles de 100. La distancia entre puntos clave estimados y de *ground-truth* no es relativa, sino absoluta.

Para resolver el problema de KLE en bases de datos donde las personas están a diferente escala (por ejemplo, SHPED), una solución, que aplicamos en el Capítulo 4, es escalar la pose estimada y la pose *ground-truth* a un mismo valor de píxeles en el torso. Con este preprocesado de datos, a la hora de usar la métrica KLE de todas las poses estimadas con el *ground-truth* de SHPED obtendremos una evaluación proporcionada.

B.3. APK

En la precisión media de puntos clave (en inglés *Average Precision of Keypoints o APK*) se asume que un punto clave está definido en base a una región espacial *ground-truth* y, por tanto, un punto clave candidato será correcto si se encuentra dentro de los píxeles en $\gamma \cdot \max(h, w)$ del punto clave del *ground-truth*, donde γ es un parámetro de umbral y h y w son la altura y la anchura de la región espacial, respectivamente (ver Figura B.1-c). Con esto se puede calcular una curva

de Precision-Recall, mostrando así también la precisión media.

La métrica de evaluación APK está propuesta por los autores Yang y Ramanan en su trabajo Mezcla de partes del cuerpo flexibles [21].

Debido a que APK es más difícil de interpretar y más lento de evaluar que otras métricas de evaluación [21], en esta tesis se ha utilizado el código fuente de APK aparecido en su trabajo 'Mezcla de partes del cuerpo flexibles' disponible en línea [76].

Una de las principales ventajas que tiene APK es que es independiente de la escala de la persona en la imagen al igual que PCP. Además, los autores de "Mezcla de partes del cuerpo flexibles"[21] exponen tres defectos de PCP (ver Sección B.1) que están corregidos en su métrica de evaluación APK. Sin embargo, el principal inconveniente de APK es que necesita, además de los puntos clave *ground-truth*, de regiones espaciales *ground-truth* cuando en PCP y KLE sólo se necesitan o de segmentos (PCP) o de puntos clave (KLE).

Bibliografía

[1] P. Viola y M. J. Jones, «Robust real-time face detection», *International Journal of Computer Vision*, vol. 57, n.º 2, págs. 137-154, mayo de 2004, ISSN: 1573-1405. DOI: 10.1023/B:VISI.0000013087.49260.fb.

[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester y D. Ramanan, «Object detection with discriminatively trained part-based models», *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 32, n.º 9, págs. 1627-1645, sep. de 2010, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2009.167.

[3] S. Zhang, *Handbook of 3D machine vision: Optical metrology and imaging*. CRC press, 2013, ISBN: 978-1-4398-7219-2.

[4] S. Mori, H. Nishida y H. Yamada, *Optical Character Recognition*. John Wiley & Sons, Inc., 1999, ISBN: 0471308196.

[5] I. Haritaoglu, D. Harwood y L. S. Davis, «W4: Real-time surveillance of people and their activities», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, n.º 8, págs. 809-830, ago. de 2000, ISSN: 0162-8828. DOI: 10.1109/34.868683.

[6] D. Joshi, Y. Rao, S. Kar, V. Kumar y R. Kumar, «Computer-vision-based approach to personal identification using finger crease pattern», *Pattern Recognition*, vol. 31, n.º 1, págs. 15-22, ene. de 1998, ISSN: 0031-3203. DOI: 10.1016/S0031-3203(97)00034-4.

[7] C. Wang, Y. Wang y A. L. Yuille, «An approach to pose-based action recognition», en *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, jun. de 2013, págs. 915-922. DOI: 10.1109/CVPR.2013.123.

[8] A. Zagouras, A. A. Argiriou, H. A. Flocas, G. Economou y S. Fotopoulos, «A machine vision based method for atmospheric circulation classification», en *International Conference on Digital Signal Processing*, jul. de 2009, págs. 1-5. DOI: 10.1109/ICDSP.2009.5201149.

[9] X. Pérez-Sala, S. Escalera, C. Ángulo y J. Gonzàlez, «A survey on model based approaches for 2d and 3d visual human pose recovery», *Sensors*, vol. 14, n.º 3, págs. 4189-4210, mar. de 2014, ISSN: 1424-8220. DOI: 10.3390/s140304189.

[10] M. Eichner, M. Marin-Jimenez, A. Zisserman y V. Ferrari, «2D articulated human pose estimation and retrieval in (almost) unconstrained still images», *International Journal of Computer Vision*, vol. 99, n.º 2, págs. 190-214, sep. de 2012, ISSN: 1573-1405. DOI: 10.1007/s11263-012-0524-9.

[11] A. Ayvaci, M. Raptis y S. Soatto, «Sparse occlusion detection with optical flow», *International Journal of Computer Vision*, vol. 97, n.º 3, págs. 322-338, mayo de 2011, ISSN: 1573-1405. DOI: 10.1007/s11263-011-0490-7.

[12] P. F. Felzenszwalb y D. P. Huttenlocher, «Pictorial structures for object recognition», *International Journal of Computer Vision (IJCV)*, vol. 61, n.º 1, págs. 55-79, ene. de 2005, ISSN: 1573-1405. DOI: 10.1023/B:VISI.0000042934.15159.49.

[13] J. Tian, Y. Lu, L. Li y W. Liu, «Tracking human poses in various scales with accurate appearance», *International Journal of Machine Learning and Cybernetics*, págs. 1-14, abr. de 2016, ISSN: 1868-808X. DOI: 10.1007/s13042-016-0537-8.

[14] J. D. Lafferty, A. McCallum y F. C. N. Pereira, «Conditional random fields: Probabilistic models for segmenting and labeling sequence data», en *Proceedings of the Eighteenth International Conference on Machine Learning (ICML)*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, págs. 282-289, ISBN: 1-55860-778-1.

[15] «People detection and tracking using stereo vision and color», *Image and Vision Computing*, vol. 25, n.º 6, págs. 995-1007, jun. de 2007, ISSN: 0262-8856. DOI: 10.1016/j.imavis.2006.07.012.

[16] D. Murray y J. J. Little, «Using real-time stereo vision for mobile robot navigation», *Autonomous Robots*, vol. 8, n.º 2, págs. 161-171, abr. de 2000, ISSN: 1573-7527. DOI: 10.1023/A:1008987612352.

[17] A. Moorhouse, A. N. Evans, G. A. Atkinson, J. Sun y M. L. Smith, «The nose on your face may not be so plain: Using the nose as a biometric», en *Crime Detection and Prevention (ICDP)*, dic. de 2009, págs. 1-6. DOI: 10.1049/ic.2009.0231.

[18] V. Ferrari, M. Marín-Jiménez y A. Zisserman, «Progressive search space reduction for human pose estimation», en *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, jun. de 2008, págs. 1-8. DOI: 10.1109/CVPR.2008.4587468.

[19] A. Cherian, J. Mairal, K. Alahari y C. Schmid, «Mixing body-part sequences for human pose estimation», en *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, jun. de 2014, págs. 2361-2368. DOI: 10.1109/CVPR.2014.302.

[20] M. I. López-Quintero, M. J. Marín-Jiménez, R. Muñoz-Salinas, F. J. Madrid-Cuevas y R. Medina-Carnicer. (2016). Stereo pictorial structure for 2d articulated human pose estimation. ver. 1.0, dirección: http://www.uco.es/investiga/grupos/ava/node/54 (visitado 21-05-2016).

[21] Y. Yang y D. Ramanan, «Articulated human detection with flexible mixtures of parts», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, n.º 12, págs. 2878-2890, dic. de 2013, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2012.261.

[22] M. Eichner y V. Ferrari, «Human pose co-estimation and applications», *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34, n.º 11, págs. 2282-2288, nov. de 2012, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2012.85.

[23] M. I. López-Quintero, M. J. Marín-Jiménez, R. Muñoz-Salinas, F. J. Madrid-Cuevas y R. Medina-Carnicer. (2016). Shped: The stereo human pose estimation dataset. ver. 1.0.1, dirección: http://www.uco.es/investiga/grupos/ava/node/47 (visitado 15-05-2016).

[24] A. Cherian, J. Mairal, K. Alahari y C. Schmid. (2014). Mixing body-part sequences for human pose estimation. ver. 1.0, dirección: http://lear.inrialpes.fr/research/posesinthewild (visitado 03-05-2016).

[25] T. Pfister, J. Charles y A. Zisserman, «Flowing convnets for human pose estimation in videos», en *2015 IEEE International Conference on Computer Vision (ICCV)*, dic. de 2015, págs. 1913-1921. DOI: 10.1109/ICCV.2015.222.

[26] «A survey of human pose estimation: The body parts parsing based methods», *Journal of Visual Communication and Image Representation*, vol. 32, págs. 10-19, oct. de 2015, ISSN: 1047-3203. DOI: 10.1016/j.jvcir.2015.06.013.

[27] M. W. Lee e I. Cohen, «8th european conference on computer vision (eccv)», en. Berlin, Heidelberg: Springer Berlin Heidelberg, mayo de 2004, cap. Human Upper Body Pose Estimation in Static Images, págs. 126-138, ISBN: 978-3-540-24671-8. DOI: 10.1007/978-3-540-24671-8_10.

[28] G. Mori, X. Ren, A. A. Efros y J. Malik, «Recovering human body configurations: Combining segmentation and recognition», en *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, vol. 2, jun. de 2004, págs. II-326-II-333. DOI: 10.1109/CVPR.2004.1315182.

[29] D. Ramanan, «Learning to parse images of articulated bodies», en *Advances in Neural Information Processing Systems 19 (NIPS)*, B. Schölkopf, J. C. Platt y T. Hoffman, eds., MIT Press, dic. de 2006, págs. 1129-1136.

[30] M. Eichner y V. Ferrari, «Better appearance models for pictorial structures», en *Proceedings of the British Machine Vision Conference (BMVC)*, BMVA Press, 2009, págs. 3.1-3.11, ISBN: 1-901725-39-1.

[31] S. Zuffi, O. Freifeld y M. J. Black, «From pictorial structures to deformable structures», en *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, jun. de 2012, págs. 3546-3553. DOI: 10.1109/CVPR.2012.6248098.

[32] M. Eichner y V. Ferrari, «European conference on computer vision (eccv)», en, K. Daniilidis, P. Maragos y N. Paragios, eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, cap. We Are Family: Joint Pose Estimation of Multiple Persons, págs. 228-242, ISBN: 978-3-642-15549-9. DOI: 10.1007/978-3-642-15549-9_17.

[33] J. J. Tompson, A. Jain, Y. LeCun y C. Bregler, «Joint training of a convolutional network and a graphical model for human pose estimation», en *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence y K. Q. Weinberger, eds., Curran Associates, Inc., dic. de 2014, págs. 1799-1807.

[34] A. Hernández, S. Sclaroff y S. Escalera, «Poselet-based contextual rescoring for human pose estimation via pictorial structures», *International Journal of Computer Vision (IJCV)*, vol. 118, n.º 1, págs. 49-64, mayo de 2016, ISSN: 1573-1405. DOI: 10.1007/s11263-015-0869-y.

[35] Y. Zhu y K. Fujimura, «8th asian conference on computer vision (accv)», en, Y. Yagi, S. B. Kang, I. S. Kweon y H. Zha, eds. Berlin, Heidelberg: Springer Berlin Heidelberg, nov. de 2007, cap. Constrained Optimization for Human Pose Estimation from Depth Sequences, págs. 408-418, ISBN: 978-3-540-76386-4. DOI: 10.1007/978-3-540-76386-4_38.

[36] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman y A. Blake, «Real-time human pose recognition in parts from single depth images», en *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, jun. de 2011, págs. 1297-1304. DOI: 10.1109/CVPR.2011.5995316.

[37] A. Baak, M. Müller, G. Bharaj, H. P. Seidel y C. Theobalt, «A data-driven approach for real-time full body pose reconstruction from a depth camera», en *International Conference on Computer Vision (ICCV)*, nov. de 2011, págs. 1092-1099. DOI: 10.1109/ICCV.2011.6126356.

[38] M. Ye, X. Wang, R. Yang, L. Ren y M. Pollefeys, «Accurate 3d pose estimation from a single depth image», en *International Conference on Computer Vision (ICCV)*, nov. de 2011, págs. 731-738. DOI: 10.1109/ICCV.2011.6126310.

[39] L. A. Schwarz, A. Mkhitaryan, D. Mateus y N. Navab, «Human skeleton tracking from depth data using geodesic distances and optical flow», *Image and Vision Computing*, vol. 30, n.º 3, págs. 217-226, mar. de 2012, ISSN: 0262-8856. DOI: 10.1016/j.imavis.2011.12.001.

[40] M. Sun, P. Kohli y J. Shotton, «Conditional regression forests for human pose estimation», en *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, jun. de 2012, págs. 3394-3401. DOI: 10.1109/CVPR.2012.6248079.

[41] G. Pons-Moll, J. Taylor, J. Shotton, A. Hertzmann y A. Fitzgibbon, «Metric regression forests for human pose estimation», en *Proceedings of the British Machine Vision Conference*, BMVA Press, sep. de 2013, págs. 4.1-4.11. DOI: 10.5244/C.27.4.

[42] M. Jiu, C. Wolf, G. Taylor y A. Baskurt, «Human body part estimation from depth images via spatially-constrained deep learning», *Pattern Recognition Letters*, vol. 50, págs. 122-129, dic. de 2014, Depth Image Analysis, ISSN: 0167-8655. DOI: 10.1016/j.patrec.2013.09.021.

[43] F. Guo y G. Qian, «Human pose inference from stereo cameras», en *Applications of Computer Vision (WACV), IEEE Winter Conference on*, feb. de 2007, pág. 37. DOI: 10.1109/WACV.2007.31.

[44] H.-D. Yang y S.-W. Lee, «Reconstruction of 3d human body pose from stereo image sequences based on top-down learning», *Pattern Recognition*, vol. 40, n.º 11, págs. 3120-3131, nov. de 2007, ISSN: 0031-3203. DOI: 10.1016/j.patcog.2007.01.033.

[45] N. D. Thang, T.-S. Kim, Y.-K. Lee y S. Lee, «Estimation of 3d human body posture via co-registration of 3d human model and sequential stereo information», *Applied Intelligence*, vol. 35, n.º 2, págs. 163-177, feb. de 2010, ISSN: 1573-7497. DOI: 10.1007/s10489-009-0209-4.

[46] G. Sheasby, J. Warrell, Y. Zhang, N. Crook y P. H. Torr, «Simultaneous human segmentation, depth and pose estimation via dual decomposition», *Proceedings of the workshop of British Machine Vision Conference (BMVC)*, sep. de 2012.

[47] J. Lallemand, M. Szczot y S. Ilic, «Articulated motion and deformable objects (amdo), 8th international conference», en, F. J. Perales y J. Santos-Victor, eds. Cham: Springer International Publishing, jul. de 2014, cap. Human Pose Estimation in Stereo Images, págs. 10-19, ISBN: 978-3-319-08849-5. DOI: 10.1007/978-3-319-08849-5_2.

[48] G. Seguin, K. Alahari, J. Sivic e I. Laptev, «Pose estimation and segmentation of multiple people in stereoscopic movies», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, n.º 8, págs. 1643-1655, ago. de 2015, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2014.2369050.

[49] A. Yao, J. Gall y L. Van Gool, «Coupled action recognition and pose estimation from multiple views», *International Journal of Computer Vision*, vol. 100, n.º 1, págs. 16-37, mayo de 2012, ISSN: 1573-1405. DOI: 10.1007/s11263-012-0532-9.

[50] J. Shen, W. Yang y Q. Liao, «Multiview human pose estimation with unconstrained motions», *Pattern Recognition Letters*, vol. 32, n.º 15, págs. 2025-2035, nov. de 2011, ISSN: 0167-8655. DOI: http://dx.doi.org/10.1016/j.patrec.2011.09.019.

[51] L. Sigal, M. Isard, H. Haussecker y M. J. Black, «Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation», *International Journal of Computer Vision (IJCV)*, vol. 98, n.º 1, págs. 15-48, sep. de 2011, ISSN: 1573-1405. DOI: 10.1007/s11263-011-0493-4.

[52] S. Amin, M. Andriluka, M. Rohrbach y B. Schiele, «Multi-view pictorial structures for 3D human pose estimation», en *Proceedings of the British Machine Vision Conference*, BMVA Press, sep. de 2013, págs. 45.1-45.12.

[53] M. Burenius, J. Sullivan y S. Carlsson, «3d pictorial structures for multiple view articulated pose estimation», en *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, jun. de 2013, págs. 3618-3625. DOI: 10.1109/CVPR.2013.464.

[54] V. Kazemi, M. Burenius, H. Azizpour y J. Sullivan, «Multi-view body part recognition with random forests», en *Proceedings of the British Machine Vision Conference (BMVC)*, BMVA Press, sep. de 2013, págs. 48.1-48.11.

[55] G. Carvalhal, R. Ferreira y A. Junqueira, «Anaglyphic three-dimensional stereoscopic printing: Revival of an old method for anatomical and surgical teaching and reporting», *Journal of Neurosurgery*, vol. 95, n.º 6, págs. 1057-1066, dic. de 2001. DOI: 10.3171/jns.2001.95.6.1057.

[56] S. J. Russell y P. Norvig, *Artificial Intelligence: A Modern Approach*, 2.ª ed. Pearson Education, 2003, ISBN: 0137903952.

[57] M. A. Fischler y R. A. Elschlager, «The representation and matching of pictorial structures», *IEEE Transactions on Computers*, vol. C-22, n.º 1, págs. 67-92, ene. de 1973, ISSN: 0018-9340. DOI: 10.1109/T-C.1973.223602.

[58] C. Rother, V. Kolmogorov y A. Blake, «Grabcut: Interactive foreground extraction using iterated graph cuts», *ACM Transactions on Graphics (SIGGRAPH)*, vol. 23, n.º 3, págs. 309-314, ago. de 2004, ISSN: 0730-0301. DOI: 10.1145/1015706.1015720.

[59] V. Ferrari, M. Marin-Jimenez y A. Zisserman, «Pose search: Retrieving people using their pose», en *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, jun. de 2009, págs. 1-8. DOI: 10.1109/CVPR.2009.5206495.

[60] P. Buehler, M. Everingham, D. Huttenlocher y A. Zisserman, «Long term arm and hand tracking for continuous sign language tv broadcasts», en *Proceedings of the British Machine Vision Conference (BMVC)*, BMVA Press, 2008, págs. 110.1-110.10, ISBN: 1-901725-36-7. DOI: 10.5244/C.22.110.

[61] M. Andriluka, S. Roth y B. Schiele, «Pictorial structures revisited: People detection and articulated pose estimation», en *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jun. de 2009, págs. 1014-1021. DOI: 10.1109/CVPR.2009.5206754.

[62] B. Sapp, A. Toshev y B. Taskar, «Cascaded models for articulated pose estimation», en *11th European Conference on Computer Vision (ECCV)*, K. Daniilidis, P. Maragos y N. Paragios, eds., Berlin, Heidelberg: Springer Berlin Heidelberg, sep. de 2010, págs. 406-420, ISBN: 978-3-642-15552-9. DOI: 10.1007/978-3-642-15552-9_30.

[63] X. Lan y D. P. Huttenlocher, «Beyond trees: Common-factor models for 2d human pose recovery», en *10th IEEE International Conference on Computer Vision (ICCV)*, vol. 1, oct. de 2005, 470-477 Vol. 1. DOI: 10.1109/ICCV.2005.48.

[64] L. Sigal y M. J. Black, «Measure locally, reason globally: Occlusion-sensitive articulated pose estimation», en *Computer Vision and Pattern Recognition (CVPR)*, vol. 2, jun. de 2006, págs. 2041-2048. DOI: 10.1109/CVPR.2006.180.

[65] S. Johnson y M. Everingham, «Clustered pose and nonlinear appearance models for human pose estimation», en *Proceedings of the British Machine Vision Conference (BMVC)*, BMVA Press, ago. de 2010, págs. 12.1-12.11, ISBN: 1-901725-40-5. DOI: 10.5244/C.24.12.

[66] P. Felzenszwalb, D. McAllester y D. Ramanan, «A discriminatively trained, multiscale, deformable part model», en *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, jun. de 2008, págs. 1-8. DOI: 10.1109/CVPR.2008.4587597.

[67] M. Eichner y V. Ferrari. (2012). Calvin upper-body detector. ver. 1.04, dirección: http://groups.inf.ed.ac.uk/calvin/calvin_upperbody_detector (visitado 18-04-2015).

[68] N. Dalal y B. Triggs, «Histograms of oriented gradients for human detection», en *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, jun. de 2005, págs. 886-893. DOI: 10.1109/CVPR.2005.177.

[69] M. Everingham, L. Gool, C. K. I. Williams, J. Winn y A. Zisserman, «The pascal visual object classes (voc) challenge», *International Journal of Computer Vision*, vol. 88, n.º 2, págs. 303-338, jun. de 2009, ISSN: 1573-1405. DOI: 10.1007/s11263-009-0275-4.

[70] R. Hartley y A. Zisserman, *Multiple View Geometry in Computer Vision*, 2.ª ed. New York, NY, USA: Cambridge University Press, 2004, ISBN: 0521540518.

[71] H. Bay, T. Tuytelaars y L. Gool, «European conference on computer vision (eccv)», Berlin, Heidelberg: Springer Berlin Heidelberg, mayo de 2006, cap. SURF: Speeded Up Robust Features, págs. 404-417, ISBN: 978-3-540-33833-8. DOI: 10.1007/11744023_32.

[72] K. Konolige, «Small vision systems: Hardware and implementation», English, en *Robotics Research*, Y. Shirai y S. Hirose, eds., Springer London, 1998, págs. 203-212, ISBN: 978-1-4471-1582-3.

[73] G. Sheasby, J. Valentin, N. Crook y P. Torr, «Asian conference on computer vision (accv)», K. M. Lee, Y. Matsushita, J. M. Rehg y Z. Hu, eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, cap. A Robust Stereo Prior for Human Segmentation, págs. 94-107, ISBN: 978-3-642-37444-9. DOI: 10.1007/978-3-642-37444-9_8.

[74] Y. Yang y D. Ramanan, «Articulated pose estimation with flexible mixtures-of-parts», en *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, jun. de 2011, págs. 1385-1392. DOI: 10.1109/CVPR.2011.5995741.

[75] M. Eichner, M. Marin-Jimenez, A. Zisserman y V. Ferrari. (2012). 2D articulated human pose estimation software. ver. 1.22, dirección: http://groups.inf.ed.ac.uk/calvin/articulated_human_pose_estimation_code (visitado 20-04-2015).

[76] Y. Yang y D. Ramanan. (2013). Articulated pose estimation with flexible mixtures of parts software. ver. 1.3, dirección: http://www.ics.uci.edu/~dramanan/software/pose (visitado 20-04-2016).

[77] S. Ren, K. He, R. Girshick y J. Sun, «Faster R-CNN: Towards real-time object detection with region proposal networks», en *Advances in Neural Information Processing Systems (NIPS) 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama y R. Garnett, eds., Curran Associates, Inc., dic. de 2015, págs. 91-99.

[78] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang y P. H. S. Torr, «Conditional random fields as recurrent neural networks», en *2015 IEEE International Conference on Computer Vision (ICCV)*, dic. de 2015, págs. 1529-1537. DOI: 10.1109/ICCV.2015.179.

[79] D. Park y D. Ramanan, «N-best maximal decoders for part models», en *Computer Vision (ICCV), 2011 IEEE International Conference on*, nov. de 2011, págs. 2627-2634. DOI: 10.1109/ICCV.2011.6126552.

[80] T. Brox y J. Malik, «Large displacement optical flow: Descriptor matching in variational motion estimation», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, n.º 3, págs. 500-513, mar. de 2011, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2010.143.

[81] B. Sapp, D. Weiss y B. Taskar, «Parsing human motion with stretchable models», en *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, jun. de 2011, págs. 1281-1288. DOI: 10.1109/CVPR.2011.5995607.

[82] M. I. López-Quintero, M. J. Marín-Jiménez, R. Muñoz-Salinas, F. J. Madrid-Cuevas y R. Medina-Carnicer, «Stereo Pictorial Structure for 2D articulated human pose estimation», *Machine Vision and Applications*, vol. 27, n.º 2, págs. 157-174, feb. de 2016, ISSN: 1432-1769. DOI: 10.1007/s00138-015-0742-6.

[83] G. Seguin, K. Alahari, J. Sivic e I. Laptev. (2015). Pose estimation and segmentation of multiple people in stereoscopic movies. ver. 1.0, dirección: http://www.di.ens.fr/willow/research/stereoseg (visitado 16-05-2016).

[84] V. Ferrari, M. Eichner, M. J. Marín-Jiménez y A. Zisserman. (2008). Buffy stickmen. ver. 3.01, dirección: http://www.robots.ox.ac.uk/~vgg/data/stickmen (visitado 13-05-2016).